Wheat and triticale whole grain near infrared hyperspectral imaging for protein, moisture and kernel hardness quantification

Sebastian Helmut Orth

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Food Science in the Faculty of AgriSciences at Stellenbosch University.

Supervisor: Prof. Marena Manley

Co-supervisor: Mr Willem Botes

Co-supervisor: Dr Paul Williams

March 2021

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Sebastian Helmut Orth

Date: 2021/02/24

Abstract

Wheat (*Triticum aestivum*) is one of the most important cereal crops grown globally. Triticale (× *Triticosecale* sp. Wittmack ex A. Camus 1927) is an important cereal crop for feed and fodder production and is also emerging as an alternative cereal for human consumption. Both these cereals are grown and produced in a diverse climatic environment and they vary with regards to their physicochemical properties. Quantitative techniques for determining protein and moisture content and kernel hardness is of importance for grading of the grains. The use of non-invasive and rapid techniques such as near-infrared hyperspectral imaging (NIR-HSI) show potential for quantification of these quality parameters. This study aimed to investigate the use of NIR-HSI (HySpex SWIR 384) with partial least squares regression (PLS-R) analysis for wheat and triticale bulk sample and single kernel image approaches.

The study considered South African wheat and triticale samples produced in three Western Cape localities, i.e. Napier, Tygerhoek and Vredenburg, comprising 180 wheat and 177 triticale samples. Of these, 39 kernels per sample were used for single kernel protein and moisture content and kernel hardness prediction, resulting in data sets with a total of 7020 wheat, 6903 triticale and 13923 combined single kernel images. This was further split into training (70%) and validation (30%) sets using the Duplex algorithm.

NIR (1100-2100 nm) hyperspectral images were acquired and the spectral data obtained for each pixel were averaged for each kernel. PLS-R was used to develop quantitative prediction models. Principal component analysis (PCA) was performed on the average spectral data and the PCA plot (PC1 vs. PC2) indicated separation between locality, with both wheat and triticale separating in the direction of PC1 from left to right. A PCA (PC1 vs. PC2) was performed for the wheat and triticale combined data set – no separation was noted. Bulk sample protein, moisture content and kernel hardness models were first evaluated which showed favourable prediction accuracy, comparable to conventional NIR spectroscopy studies performed on wheat and triticale. The combined wheat and triticale data sets for protein and moisture content and kernel hardness prediction had RMSEP-values of 0.41%, 0.49% and 8.66, respectively.

Single kernel analysis involved two main quantitative data analysis methods (PLS-R and Robust-PLS) which were tested with an independent test set. The results being favourable for the conventional PLS-R method when only the validation set RMSEP (protein content: 0.37-0.84%, moisture content: 0.23-0.57% and kernel hardness: 1.74-3.64) was considered. The independent test set for protein content prediction achieved better results with the Robust-PLS (RMSEP protein content: 1.95-2.37%) method, proving that the method did indeed have an effect on making the calibration data sets more robust.

Spectral imaging showed that it is capable to accurately quantifying protein and moisture content and kernel hardness of bulk and single kernel samples – good robust models proved to optimally quantify these parameters. The technique shows good potential for further study and to build onto the current data sets in order to increase variance across seasons. Further the technique showcases the functionality of SK NIR-HSI analysis and can be used both as a quality control measure and as an early generation selection method by the grain breeding sector.

Opsomming

Koring (*Triticum aestivum*) is een van die wêreld se belangrikste graan gewasse. Korog (× *Triticosecale* sp. Wittmack ex A. Camus 1927) is ʼn belangrike graan gewas vir aangeplante weiding en kuilvoer produksie en is ook ʼn opkomende alternatiewe graan vir menslike gebruik. Albei hierdie graan soorte word in ʼn diverse klimatologiese omgewing verbou en daar is ʼn groot variasie tussen grootmaat monsters en tussen enkel sade vanuit ʼn monster. Kwantitatiewe tegnieke om graan proteïen- en voginhoud en hardheid te bepaal is van belang vir die gradering daarvan. Die gebruik van nie-indringende en vinnige tegnieke soos naby infrarooi (NIR) hiperspektrale beelding wys potensiaal vir kwantifisering van kwaliteiteienskappe. Hierdie studie was daarop gemik om ondersoek in te stel tot die gebruik van NIR hiperspektrale (HySpex SWIR 384) beelding met parsiële kleinste kwadrate regressie as die data analise metode, vir koring en ook korog monsters op ʼn grootmaat monster asook ʼn enkel saad beelding benadering.

Die studie het Suid-Afrikaanse koring en korog monsters oorweeg wat verbou is in drie distrikte in die Wes-Kaap provinsie naamlik Napier, Tygerhoek en Vredenburg, wat verder afgebreek is na 180 koring en 177 korog monsters. Vanaf die grootmaat monsters is 39 sade per monster gebruik vir enkel saad analise vir proteïen, vog en hardheid inhoud bepalings, wat ʼn totaal van 7020 koring, 6903 korog en ʼn gekombineerde 13923 sade opmaak vir elke data stel.

NIR hiperspektrale beelding (1100-2100nm) is gebruik om pixel en daaropvolgende spektrale data te verkry vanaf die sade en parsiële kleinste kwadrate regressie is gebruik as die kwantitatiewe data analise metode. Hoofkomponent analise (HKA) vir HK1 teen HK2 is uitgeoefen vir die bepaling van skeiding tussen monsters gebaseer op verbouings lokaliteit. Beide koring en korog datastelle wys daarop dat daar skeiding oor HK1 is van links na regs. ʼn HKA (HK1 teen HK2) is ook toegepas op die kombinasie datastel vir koring en korog, dit het geen skeiding tussen die twee graan soorte getoon nie. Grootmaat proteïen, vog en korrel hardheid modelle is toegepas op koring en korog en het gewys op gunstige voorspellings akkuraatheid wat vergelykbaar is met studies wat gefokus het op die gebruik van konvensionele NIR spektroskopie op koring en korog. Die gekombineerde data stelle vir proteïen- en voginhoud en hardheid bepaling het ʼn gemiddelde vierkantswortel fout van voorspelling (GVFV) waardes van 0.41%, 0.49% en 8.66, onderskeidelik gehad.

Vir enkel saad analise, is twee kwantitatiewe data analise metodes gebruik (parsiële kleinste kwadraat regressie en robuust parsiële kleinste kwadraat regressie) wat getoets is teenoor ŉ onafhanklike toets stel. Die resultate was gunstig vir die konvensionele parsiële kleinste kwadraat regressie metode wanneer slegs gekyk is na die GVFV van die validasie stel. Die onafhanklike toets stel vir proteïeninhoud bepaling het ŉ beter GVFV gehad vir die robuust parsiële kleinste kwadrate regressie en wys daarop dat die kalibrasie van die modelle meer robuuste voorspellings maak.

Spektrale beelding het gewys dat dit ŉ akkurate metode is om proteïen- en voginhoud en hardheid van grootmaat sowel as enkel sade te bepaal. Met optimale resultate geskik vir meer robuuste modelle vir verdere kwantifisering van kalibrasie parameters. Die tegnieke wys potensiaal vir verdere studie en om verder te bou op die huidige data stelle vir meer variasie oor seisoene. Verder word die funksionaliteit van NIR hiperspektrale beelding uit gewys en die metode kan sy plek vind in kwaliteit beheer sowel as graan seleksie in die graan teler sektor.

Acknowledgements

My utmost gratitude to my supervisors Prof. Manley, Dr Williams and Mr Willem Botes for your continuous support of my research. Prof. Manley your patience, motivation and moral support were greatly appreciated throughout the research and thesis writing phases. The guidance has truly made me a more rounded and accomplished researcher.

Dr Stefan Hayward and Dr Timo Tait your continued support and motivation within the lab and the readiness to always go and evaluate some beer when things became a little bit “rough” is something I will definitely cherish. The deep and focussed conversations regarding the scientific domain has allowed me to suck up knowledge somewhat like a sponge.

The department of Food Science for allocating the departmental bursary and Prof Manley for helping me out from her research funds, for this I am eternally grateful.

Prof. Gunnar Sigge and Prof. Pieter Gouws your continued support and levelling conversations were a weekly highlight – keep it up.

The Stellenbosch University Plant Breeding Laboratory for providing me with samples so that I could perform my study.

The Winter Cereal Trust for providing project funding (WCT/W/2020/03)

To my friends, Joachim, Janus and Francois for always being up for an excursion and a braai – dit maak dit LUUKS.

My father Dr Claus Orth for continued support, even if we bump heads on occasion I could not have done this without you.

Table of contents

Abstract ii

Opsomming iv

Acknowledgements vi

Table of contents vii

List of Figures ix

List of Tables xii

List of Abbreviations xiii

Chapter 1: Introduction 1

The aim of this study was to: 4

The specific objectives was to: 4

References 5

Chapter 2: Literature review 7

2.1 Introduction 7

2.2 Wheat and triticale 8

2.2.1 Kernel morphology 9

2.2.2 Pericarp 10

2.2.3 Endosperm 11

2.2.4 Embryo 12

2.3. Protein in small grains 12

2.4 Protein content determination in wheat 13

2.5. Endosperm texture and kernel hardness 13

2.5.1 Kernel hardness determination methods 15

2.5.2 Single Kernel Characterisation System (SKCS) 16

2.5.3 Pohl farinator 17

2.5.4 Particle size index (PSI) 17

2.5.5 Near-infrared spectroscopy – kernel hardness 18

2.6 Near-infrared spectroscopy 18

2.6.1 Fundamental principles of near-infrared spectroscopy 19

2.6.2 Chemometrics 21

2.7 Fundamentals of hyperspectral imaging 21

2.7.1 Hyperspectral image acquisition 22

2.8 Near-infrared spectroscopy and hyperspectral imaging of small grains 23

2.8.1 Conventional NIR spectroscopy 23

2.8.2 NIR hyperspectral imaging 26

2.9 Conclusion 28

2.10 References 29

Chapter 3: Materials and methods ... 35
3.1 Wheat and triticale samples ... 35
3.2 Protein and moisture content and kernel hardness determination ... 35
3.3 Near-infrared hyperspectral image system setup and image acquisition ... 35
3.4 Hyperspectral image analysis ... 37
3.5 Partial least squares regression ... 37
3.6 References ... 40
Chapter 4: Results and discussion ... 41
4.1 Exploratory analysis ... 41
4.2 Bulk wheat and triticale PLS regression models ... 45
4.2.1 Protein content ... 45
4.2.2 Moisture content ... 49
4.2.3 Kernel hardness ... 53
4.3 Single kernel prediction models ... 58
4.3.1 Single kernel protein ... 59
4.4 Single kernel moisture content ... 72
4.5 Single kernel hardness ... 83
4.6 Single kernel protein content independent test set ... 92
4.8 Conclusion ... 98
4.9 References ... 99
Chapter 5: General discussion and conclusions ... 100
References ... 105

List of Figures

Figure 2.1 NIR spectroscopy and classic NIR-HSI sensing modes including reflectance, transmittance and interactance modes of detection.

Figure 3.1 Near-infrared hyperspectral imaging setup with the HySpex SWIR 384 camera equipped with a HgCdTe sensor, two halogen light sources and a translation stage.

Figure 3.2 The black nylon tray used for sample presentation for spectral imaging allowing seven sets of 39 single kernels per sample to be imaged simultaneously.

Figure 3.3 A flow diagram summarising the methodology used to build and evaluate models for wheat and triticale protein and moisture content and also kernel hardness prediction.

Figure 4.1 Whole kernel, raw average spectra of the (a) wheat, (b) triticale and (c) wheat and triticale combined data sets. PCA plots of PC1 vs. PC2 for the corresponding mean-centred spectral data with score values coloured based on location (Napier, Tygerhoek and Vredenburg) for (d) wheat, (e) triticale and on the (f) type of grain (wheat and triticale) in the combined data set.

Figure 4.2. PCA plot of PC1 vs. PC2 of spectra from the combined data set pre-treated with SNV, DT and 2nd derivative, showing distinct clustering of wheat samples from an experimental plot mainly from Tygerhoek.

Figure 4.3 Histograms illustrating distribution of reference data, i.e. protein content for (a) wheat, (b) triticale and (c) wheat and triticale combined and moisture content for (c) wheat, (d) triticale and (e) wheat and triticale combined.

Figure 4.4 (a) Average pre-processed spectra of the combined data set for protein content prediction model 1 (Table 4.2) , (b) variable importance in projection and (c) latent variables for LV 1, LV 2 and LV 11.

Figure 4.5 Latent variables vs. standard error of cross-validation, -calibration and -prediction for the combined data set protein prediction model 1 (Table 4.2.).

Figure 4.6 Measured vs. predicted protein content for PLS regression model 1 with SNV, DT and 2nd derivative pre-treatment for the combined wheat and triticale data set using 11 LV's and 249 samples in the training set and 106 in the validation set.

Figure 4.7. (a) Average pre-processed spectra of the combined data set for moisture content prediction model (Table 4.3), (b) variable importance in projection and (c) latent variables for LV 1, LV 2 and LV 11.

Figure 4.8 Latent variables vs. standard error of cross-validation, -calibration and -prediction error for the combined data set moisture prediction model 6 (Table 4.3).

Figure 4.9 Measured vs. predicted moisture content for PLS regression model 6, with SNV DT and 2nd derivative pre-treatment for the combined wheat and triticale data set using 12 LV's and 249 samples in the training set and 107 in the validation set.

Figure 4.10 Histograms illustrating distribution of kernel hardness reference data for (a) wheat, (b) triticale and (c) the combined data sets**.**

Figure 4.11. (a) Average pre-processed spectra of the combined data set for kernel hardness prediction model 1 (Table 4.4), (b) variable importance in projection and (c) LV's for (LV 1, LV 2 and LV 11).

Figure 4.12 Latent variables vs standard error of cross-validation, -calibration and -prediction error for the combined data set hardness prediction model 1 (Table 4.4).

Figure 4.13 Measured vs. predicted kernel hardness values for the PLS regression model with SNV and 2nd derivative pre-treatment for the combined data set using 7 LV's and a calibration set of 234 samples and a validation set of 100.

Figure 4.14 (a) Pre-processed SK spectra of the wheat data set for protein content prediction model 3 (Table 4.6), (b) variable importance in projection vs wavebands and (c) latent variables for LV 1, LV 2 and LV 18.

Figure 4.15. Latent variables vs standard error of cross-validation, -calibration and -prediction error for wheat protein prediction model 3 (Table 4.6).

Figure 4.16 Measured vs. predicted protein content for PLS regression model 3 for the wheat data set pre-treated with SNV (Table 4.6) using 18 LV's and a calibration set of 2883 and validation set of 1148 SK's.

Figure 4.17 (a) Pre-processed spectra of the SK triticale data set for protein content prediction model 2 (Table 4.6), (b) variable importance in projection plot and (c) LV's plot for LV 1, LV 2 and LV 18.

Figure 4.18 Latent variables vs standard error of cross-validation, -calibration and -prediction error for SK triticale protein content prediction model 2 (Table 4.6).

Figure 4.19 Measured vs. predicted protein content for PLS regression model 2 for SNV and DT pre-treated triticale SK data set (Table 4.6) with 18 LV's and a calibration set of 2940 and a validation set of 1208 SK's.

Figure 4.20. (a) Pre-processed spectra of the combined SK data set for protein content prediction model 2 (Table 4.6), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 18.

Figure 4.21 Latent variables vs standard error of cross-validation, -calibration and -prediction error for wheat and triticale protein prediction model 2 (Table 4.6).

Figure 4.22 Measured vs. predicted protein content for PLS regression model with SNV and DT pre-treatment for the combined wheat and triticale SK data set (Table 4.6) using 18 LV's and a calibration set of 6459 and 2669 SK's.

Figure 4.23 (a) Pre-processed spectra of the wheat SK data set for moisture content prediction model 3 (Table 4.8) wheat moisture content prediction, (b) variable importance in projection plot and (c) LV's for LV 1, LV 2 and LV.

Figure 4.24 Latent variables vs standard error of cross-validation, -calibration and -prediction error for SK wheat moisture content prediction model 3 (Table 4.8).

Figure 4.25 Measured vs. predicted moisture content for PLS regression model with SNV pre-treatment for the SK wheat data set model 3 (Table 4.8) using 16 LV's and a calibration set of 3803 and a validation set of 1630 SK's.

Figure 4.26 (a) Pre-processed spectra of the triticale SK data set for moisture content prediction for model 3 (Table 4.8), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 14.

Figure 4.27 Latent variables vs standard error of cross validation, -calibration and -prediction for SK triticale moisture content prediction model 3 (Table 4.8).

Figure 4.28 Measured vs. predicted moisture content for PLS regression model with SNV pre-treatment for the SK triticale data set for model 3 (Table 4.8) using 14 LV's and a calibration set of 3296 and validation set of 1325 SK's.

Figure 4.29 (a) Pre-processed spectra of the combined SK data set for moisture content prediction model 3 (Table 4.8)), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 17.

Figure 4.30. Latent variables vs standard error of cross-validation, -calibration and -prediction error for SK combined wheat and triticale moisture content prediction model 3 (Table 4.8).

Figure 4.31 Measured vs. predicted moisture content for PLS regression model with SNV pre-treatment for the SK combined wheat and triticale data set (Table 4.8) using 17 LV's and a calibration set of 6722 and validation set of 2894 SK's.

Figure 4.32 (a) Pre-processed spectra of the wheat SK data set for kernel hardness prediction model 2 (Table 4.9), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 3.

Figure 4.33 Latent variables vs standard error of cross-validation, -calibration and -prediction for SK wheat hardness prediction model 2 (Table 4.9).

Figure 4.34 Measured vs. predicted hardness for PLS regression model 2 with GLS and SNV pre-treatment for the SK wheat data set (Table 4.9) with 3 LV's and a calibration set of 2362 and validation set of 886 SK's.

Figure 4.35 (a) Pre-processed spectra of the triticale SK data set for kernel hardness prediction model 2 (Table 4.9), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 3.

Figure 4.36 Latent variables vs standard error of cross-validation, -calibration and -prediction for hardness prediction model 2 (Table 4.9).

Figure 4.37 Measured vs. predicted hardness for PLS regression model with GLS and SNV spectral pre-treatment for the SK triticale data set model 2 (Table 4.9) using 4 LV's and a calibration set consisting of 1640 and validation of 658 single kernels.

Figure 4.38 (a) Pre-processed of the combined SK data set for kernel hardness prediction model 2 (Table 4.9), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 3.

Figure 4.39. Latent variables vs standard error of cross-validation, -calibration and -prediction for hardness prediction model 2 (Table 4.9).

Figure 4.40 Measured vs. predicted hardness for PLS regression model with GLS and SNV spectral pre-treatment for the SK combined wheat and triticale data set model 2 (Table 4.9) using 4 LV's and a calibration set of 3259 and validation set of 1288 single kernels.

Figure 4.41 Predicted protein content vs. residual reference values for wheat SK test set, obtained for conventional (a) PLS and (b) Robust-PLS methods (Tables 4.5 and 4.6).

Figure 4.42 Predicted protein content vs. residual reference values for triticale SK test set, obtained for conventional (a) PLS and (b) Robust-PLS methods (Tables 4.5 and 4.6).

Figure 4.43 Predicted protein content vs. residual reference values for the combined SK test set, obtained for conventional (a) PLS and (b) Robust-PLS (Tables 4.5 and 4.6).

Figure 4.44 Showing R^2 plotted against RPD statistics for the SK wheat, triticale and combined data set models.

List of Tables

Table 2.1 Hardness index categories for soft to hard kernels as adapted from AACC International method 55-31.01 (AACC Approved Methods of Analysis, 1999a)

Table 2.3 Differences between conventional NIR spectroscopy, hyperspectral imaging, RGB imaging and multispectral imaging (adapted from Wu and Sun, 2013**)**

Table 3.1 Number of samples in training and validation sets for bulk and single kernel wheat, triticale and combined data sets selected using the DUPLEX algorithm

Table 4.1 Descriptive statistics for protein and moisture content (%) of wheat and triticale samples

Table 4.2 Calibration and validation statistics for protein content PLS regression models for bulk wheat, triticale and combined data sets using different pre-processing methods. The best prediction based on lowest RMSEP is indicated in bold

Table 1.3 Calibration and validation statistics for predicted moisture PLS regression models for bulk wheat, triticale and combined data sets using different pre-processing methods. The best prediction based on lowest RMSEP is indicated in bold

Table 4.4 Calibration and validation statistics for predicted kernel hardness PLS regression models for bulk wheat, triticale and combined data sets using different pre-processing methods. The best prediction based on lowest RMSEP is indicated in bold

Table 4.5 Calibration and validation statistics for predicted protein content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and the robust-PLS outlier removal method. The best prediction based on lowest RMSEP is indicated in bold

Table 4.6 Calibration and validation statistics for predicted protein content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and removing outliers manually. The best prediction based on lowest RMSEP is indicated in bold

Table 4.7 Calibration and validation statistics for predicted moisture content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and the robust-PLS outlier removal method. The best prediction based on lowest RMSEP is indicated in bold

Table 4.8 Calibration and validation statistics for predicted moisture content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and removing outliers manually. The best prediction based on lowest RMSEP is indicated in bold

Table 4.9 Calibration and validation statistics for predicted hardness PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and removing outliers manually. The best prediction based on lowest RMSEP is indicated in bold

Table 4.10 RMSEP for SK independent test set of wheat, triticale, and the combined data set tested against the best prediction models for the robust-PLS and conventional PLS methods (Tables 4.5 and 4.6)

List of Abbreviations

ANN	Artificial neural networks
CA	Cluster analysis
CV	Cross validation
DA	Discriminate analysis
DC	Direct current
DT	Detrend
FA	Factor analysis
FT	Fourier transform
GLS	Generalised least squares
HI	Hardness index
HSI	Hyperspectral imaging
IR	Infrared
LDA	linear discriminant analysis
LV	Latent variables
MC	Mean centering
MIR	Mid-infrared
MLR	Multiple linear regression
NIR	Near-infrared
NIRS	Near-infrared spectroscopy
NN	Neural networks
OSC	Orthogonal signal correction
SU-PBL	Stellenbosch University Plant Breeding Laboratory
PC	Principal component
PCA	Principal component analysis
PCR	Principal component regression
PLSR	Partial least squares regression
PRI	Pearling resistance index
PSI	Particle size index
PTFE	Polytetrafluoroethylene
RGB	Red Green Blue
RPD	Ratio of performance to deviation
RMSEC	Root mean square error of calibration
RMSECV	Root mean square error of cross-validation
RMSEP	Root mean square error of prediction
RSIMPLS	Robust-PLS
SD	Standard deviation
SDS-PAGE	Sodium dodecyl-sulphate polyacrylamide gel electrophoresis

SECV	Standard error cross validation
SEP	Standard error prediction
SIMPLS	Straightforward Implementation of a statistically inspired Modification of the PLS method
SK	Single kernel
SKCS	Single Kernel Characterisation System
SNV	Standard normal variate
SWIR	Short wave infrared
UK	United Kingdom
VIP	Variable importance in projection
WHI	Wheat hardness index
XT	Training set
XV	Validation set

Chapter 1: Introduction

Commercial wheat, used for food applications, consists of two primary species, i.e. bread (*Triticum aestivum*) and durum (*Triticum turgidum*) wheat. In contrast to this triticale (× Triticosecale sp. Wittmack ex A. Camus 1927) is an intergeneric hybrid of wheat (*Triticum* spp.) and rye (*Secale cereale*). Triticale usage is mostly as feed and fodder and shows some potential to be used as a food source for human consumption (Mcgoverin et al., 2011; Zhu 2018). The grading of wheat (and for practical reasons triticale) is based on its chemical and physical properties, i.e. protein content, moisture content and also physically by hectolitre mass determination. This ultimately affects wheat and triticale final application, nutritional impact and the commercial value. Producers and graders consider government defined grading parameters for acquiring the best grain grade for optimum economic gain. Whereas millers require optimum flour yield as consistent baking performance is of importance to bakers for the production of consumer acceptable products. The optimisation of planting material by plant breeders is of constant focus, with specific traits being exploited that influence biotic and abiotic stress tolerance, grain yield and quality parameters.

Protein content is a major quality indicator in cereals that are milled for flour. More specifically gluten proteins have an impact on dough forming and rheological properties. Directly influencing the overall baking and dough proofing aspects related to breadmaking and other baked goods (Shewry *et al.*, 2002; Uthayakumaran and Stoddard, 1999). Furthermore, proteins are part of the structural aspects of grain kernels and are important aspect regarding kernel hardness (Stenvert and Kingswood, 1977). Analysis techniques which are destructive in nature such as the Dumas combustion and Kjeldahl digestion methods are often used for protein content determination.

Moisture content influences storage conditions and energy input of wheat towards drying the grain in silos. The moisture content of wheat also influences milling performance and standardisation of wet milling procedures. And finally it will influence the shelf life of the final product, i.e. flour (Sun and Woods, 1993). Analysis techniques are often time consuming and destructive in nature and involve oven drying and milling of the whole grain into flour.

Wheat endosperm texture is of great importance to the milling industry. Hard grains (vitreous endosperm) give a course flour with high levels of starch damage and soft (opaque) kernels produce a soft flour with low levels of starch damage (Bolling, 1987). Wheat hardness and its chemical makeup has been studied extensively (Cobb 1896; Miller et al., 1981; Pomeranz et al., 1984; Greenwell and Schofield 1986; Greenwell and Schofield 1989; Pomeranz and Williams 1990; Bechtel et al., 1996; Dowell 2000; Turnbull and Rahman 2002; Turnbull et al., 2003). Wheat hardness content is classically determined by destructive measures such as particle size index and the single kernel characterisation system.

Non-destructive conventional NIR spectroscopy and NIR hyperspectral imaging (NIR-HSI) methods are available and has been extensively researched for whole wheat analysis (Delwiche and Hruschka, 2000; Maghirang and Dowell, 2003; Igne *et al.*, 2007; Manley *et al.*, 2013; Mahesh *et al.*, 2014a), however, these measurements are typically done on bulk samples. Single kernel analysis (Delwiche, 1993; Delwiche, 1995, 1998; Nielsen *et al.*, 2003; Armstrong *et al.*, 2006; Bramble *et al.*, 2006; Caporaso *et al.*, 2018) can add value to breeding programmes especially for analysis of early generation material. This will enable the selection of single kernels with specific traits. And ultimately single kernel analysis gives a good indication of the distribution of quality parameters, i.e. protein, moisture and hardness content within a sample

Similar to NIR spectroscopy, NIR-HSI shares the common advantages of being non-invasive, rapid and non-destructive. Once a model is established using NIR-HSI, multiple kernels can be imaged simultaneously and provide results on a single kernel basis. NIR hyperspectral imaging thus poses a powerful tool for routine analysis and for providing information not viable with conventional analytical techniques (Manley, 2014). This allows for characterisation of single kernels of wheat based on its, protein, moisture and grain hardness content. Subsequently enabling wheat breeders to make calculated pre- and early generation selection of their grain seeds before propagation commences. Due to the non-invasive nature of NIR technology it is possible to perform analysis on sensitive materials without having to perform any sample preparation or destroying the sample (Fox and Manley, 2014).

NIR-HSI is a non-destructive, rapid and unbiased technique, that utilises the fundamentals of spectroscopy and imaging, enabling multidimensional spectral and spatial information to be acquired simultaneously (ElMasry et al., 2012; Feng and Sun 2012). The application of hyperspectral imaging on bulk and single kernel wheat has been extensively researched and it involves both qualitative and quantitative studies. Studies specifically on single kernel analysis using conventional Near-infrared spectroscopy and hyperspectral imaging has been reviewed by Fox and Manley, (2014). Further and updated methods have been reviewed in Chapter 2 of this manuscript, bringing attention to the shortfalls of the studies.

Delwiche and Hruschka (2000) used near-infrared (NIR) reflectance spectroscopy to estimate bulk sample protein from single kernel spectral readings and showed that an increase in sample size (10 – 100) resulted in a decrease in standard error of cross-validation (SECV; 0.385 – 0.162%). Protein content prediction studies using NIR spectroscopy done for triticale by Fontaine *et al* (2002) resulted in similar prediction accuracies (SECV = 0.235%; R^2 of 0.98). Igne *et al* (2007) also used NIR spectroscopy for the prediction of whole-grain triticale moisture (SEP = 0.29%) and protein (SEP = 0.30%) content. These authors showed that prediction models developed for wheat were appropriate for triticale protein prediction (SEP = 0.38%), and also for moisture prediction (SEP = 0.37%). In 2013, Manley *et al*. (2013) developed NIR spectroscopy calibrations for whole grain triticale quality parameter predictions showing whole grain (SEP = 0.67%; R^2 = 0.92) calibrations to be less accurate than that of ground grain (SEP = 0.52%; R^2 = 0.95). The authors represented thus far all showed work done on small and limited datasets which resulted in limited and proof of concept models.

The use of NIR-HSI for protein prediction of bulk Canadian whole grain wheat has been explored (Mahesh *et al.*, 2014). The authors found that partial least squares (PLS) regression (SEP = 1.76%; R^2 = 0.46) gave better results than principal component (PCR) regression (SEP = 2.02%; R^2 = 0.38). Single kernel PLS-R modelling for protein content prediction using NIR-HSI as the analytical tool has been explored by Caporaso et al., 2018. In their study a large dataset of 3250 for the calibration and 868 kernels for the validation set were used. The authors obtained a route mean square error (RMSE) of 0.86 with an R^2 of 0.82 for the calibration set and a RMSE of 0.94 with an R^2 of 0.79. Considering the work done previously in the field, one of the greatest shortfalls was researchers not allowing for enough sample variance, but rather creating their own variance through moisture content adjustment.

NIR-HSI allows for single kernels of different breeding lines to be simultaneously analysed to obtain intrinsic spectral information of the samples. Using this technique as a non-subjective and non-invasive method, models can be used to rapidly quantify for protein, moisture and hardness content. Subsequently this will benefit the grain farmers and millers by providing rapid information towards quantified quality parameters of wheat and triticale single kernels. Not only can this provide for time and cost saving within the breeding sector, it can also offer a rapid non-destructive grading technique at mills and silos.

The aim of this study was to:

investigate wheat and triticale NIR-HSI partial least squares regression models to accurately predict protein and moisture content and kernel hardness.

The specific objectives were to:

1. develop NIR-HSI calibration models for bulk sample quantification of protein and moisture content as well as kernel hardness; and
2. develop NIR HSI calibration models for single kernel quantification of protein and moisture content as well as kernel hardness.

References

Armstrong, P.R., Maghirang, E.B., Xie, F. & Dowell, F.E. (2006). Comparison of dispersive and Fourier-transform NIR instruments for measuring grain and flour attributes. *Applied Engineering in Agriculture*, **22**, 453-457.

Bechtel, D.B., Wilson, J.D. & Martin, C.R. (1996). Determining endosperm texture of developing hard and soft red winter wheats dried by different methods using the single-kernel wheat characterization system. *Cereal Chemistry*, **73**, 567-570.

Bolling, H. (1987). Milling quality of wheat. *European Conference on Food Science and Technology*, 259–284.

Bramble, T., Dowell, F.E. & Herrman, T.J. (2006). Single-kernel near-infrared protein prediction and the role of kernel weight in hard red winter wheat. *Applied Engineering in Agriculture*, **22**, 945-949.

Caporaso, N., Whitworth, M.B. & Fisk, I.D. (2018). Protein content prediction in single wheat kernels using hyperspectral imaging. *Food Chemistry*, **240**, 32–42.

Cobb, N.A. (1896). The hardness of the grain in the principal varieties of wheat. *Agric. Gazette*, 279–298.

Delwiche, S.R. (1995). Single wheat kernel analysis by near-infrared transmittance: protein content. *Cereal Chemistry*, **72**, 11-16.

Delwiche, S.R. (1998). Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. *Journal of Cereal Science*¸ **27**, 241-254.

Delwiche, S.R. & Hruschka, W.R. (2000). Protein content of bulk wheat from near-infrared reflectance of individual kernels. *Cereal Chemistry*, **77**, 86-88.

Dowell, F.E. (2000). Differentiating vitreous and nonvitreous durum wheat kernels by using near-infrared spectroscopy. *Cereal Chemistry*, **77**, 155-158.

ElMasry, G., Kamruzzaman, M., Sun, D.W. & Allen, P. (2012). Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review. *Critical Reviews in Food Science and Nutrition*, **52**, 999–1023.

Feng, Y.-Z. & Sun, D.-W. (2012). Application of hyperspectral imaging in food safety inspection and control: a review. *Critical Reviews in Food Science and Nutrition*, **52**, 1039–1058.

Fox, G. & Manley, M. (2014a). Applications of single kernel conventional and hyperspectral imaging near infrared spectroscopy in cereals. *Journal of the Science of Food and Agriculture*, **94**, 174–179.

Greenwell, P. & Schofield, J. (1986). A starch granule protein associated with endosperm softness in wheat. *Cereal chemistry*, **63**, 379–380.

Greenwell, P. & Schofield, J.D. (1989). The chemical basis of grain hardness and softness. *H. Salovaara (Ed.) Wheat End-Use Properties, Proceedings ICC '89 Symposium (Lahti, Finland)*, 59–72.

Igne, B., Gibson, L.R., Rippke, G.R., Schwarte, A. & Hurburgh, C.R. (2007). Triticale moisture and protein content prediction by near-infrared spectroscopy (NIRS). *Cereal Chemistry*, **84**, 328–330.

Maghirang, E.B. & Dowell, F.E. (2003). Hardness measurement of bulk wheat by single-kernel visible and near-infrared reflectance spectroscopy. *Cereal Chemistry*, **80**, 316-322.

Mahesh, S., Jayas, D.S., Paliwal, J. & White, N.D.G. (2014a). Comparison of partial least squares regression (plsr) and principal components regression (pcr) methods for protein and hardness predictions using the near-infrared (nir) hyperspectral images of bulk samples of canadian wheat. *Food and Bioprocess Technology*, **8**, 31–40.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. *Chemical Society Reviews*, **43**, 8200–8214.

Manley, M., McGoverin, C.M., Snyders, F., Muller, N., Botes, W.C. & Fox, G.P. (2013). Prediction of triticale grain quality properties, based on both chemical and indirectly measured reference methods, using near-infrared spectroscopy. *Cereal Chemistry*, **90**, 540–545.

Mcgoverin, C.M., Snyders, F., Muller, N., Botes, W., Fox, G. & Manley, M. (2011). A review of triticale uses and the effect of growth environment on grain quality. *Journal of the Science of Food and Agriculture*, **91**, 1155–1165.

Miller, B.S., Afework, S., Hughes, J.W. & Pomeranz, Y. (1981). Wheat hardness: time required to grind wheat with the brabender automatic. micro hardness tester. *Journal of Food Science*, **46**, 1863-1865.

Nielsen, J.P., Pedersen, D.K. & Munck, L. (2003). Development of nondestructive screening methods for single kernel characterization of wheat. *Cereal Chemistry*, **80**, 274-280.

Pomeranz, Y., Bolling, H. & Zwingelberg, H. (1984). Wheat hardness and baking properties of wheat flours. *Journal of Cereal Science*, **2**, 137–143.

Pomeranz, Y. & Williams, P.C. (1990). Wheat hardness. Its genetic, structural, and biochemical background, measurement, and significance. *Advances in Cereal Science and Technology*, **10**, 471–544.

Delwiche, S.R. (1993). Measurement of single-kernel wheat hardness using near-infrared transmittance. *Transactions of the ASAE*, **36**, 1431–1437.

Shewry, P.R., Halford, N.G., Belton, P.S. & Tatham, A.S. (2002). The structure and properties of gluten : an elastic protein from wheat grain. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **357**, 133–142.

Stenvert, N.L. & Kingswood, K. (1977). The influence of the physical structure of the protein matrix on wheat hardness. *Journal of the Science of Food and Agriculture*, **28**, 11–19.

Sun, D.W. & Woods, J.L. (1993). The moisture content/relative humidity equilibrium relationship of wheat - a review. *drying technology*, **11**, 1523–1551.

Turnbull, K.M., Marion, D., Gaborit, T., Appels, R. & Rahman, S. (2003). Early expression of grain hardness in the developing wheat endosperm. *Planta*, **216**, 699–706.

Turnbull, K.M. & Rahman, S. (2002). Endosperm texture in wheat. *Journal of Cereal Science*, **36**, 327–337.

Uthayakumaran, S. & Stoddard, F.L. (1999). Effect of varying protein content and glutenin-to-gliadin ratio on the functional properties of wheat dough. *Cereal Chemistry*, **76**, 389-394.

Zhu, F. (2018). Triticale: Nutritional compositian and food uses. *Food Chemistry*, **241**, 468–479.

Chapter 2: Literature review

2.1 Introduction

Various food products are made globally using wheat and its flour derivatives. The two species accounting for the majority of consumption are bread wheat (*Triticum aestivum)* and durum wheat (*Triticum turgidum L. ssp. durum)*. The primary use of triticale (× Triticosecale sp. Wittmack ex A. Camus 1927), on the other hand, is as livestock (chickens, pigs, geese, cattle, and sheep) feed, where it is used in all its forms, i.e. grain, forage, silage, hay and straw (McGoverin *et al.*, 2011). However, as the world's population becomes ever more health conscious, the need for alternative cereal grains is on the increase. Triticale is thus now seen to be increasingly used to produce food products such as pasta, bread, tortillas, biscuits and yogurt (Zhu, 2018). It is also used to produce edible films, malt and it is used in the spirits industry.

Near-infrared (NIR) spectroscopy is routinely used during plant breeding and in grain industries for the prediction of physicochemical properties (Williams *et al*., 2019). The most common industrial applications include prediction of constituents such as protein and moisture, both of which are strong absorbers in the NIR spectral region. More recently, NIR hyperspectral imaging (NIR-HSI) has also become recognised as a non-destructive and non-invasive technique for the quantitative and qualitative analysis of cereal grains (Sendin *et al.*, 2018). NIR-HSI has the added advantage of a spatial dimension making it suitable for heterogeneous samples (Manley, 2014) or e.g. simultaneous analysis of multiple single cereal grains. Wheat and triticale breeding programmes often deal with small sample sizes (*ca*. 5 g) and fast, non-destructive analysis of properties such as protein and moisture content as well as kernel hardness is essential for efficient breeding practices as highlighted in Chapter 1.

In this review the morphology of wheat and triticale are considered with specific reference to genetic differences, protein content and distribution and kernel hardness. Conventional methods routinely used to measure kernel hardness are reviewed. The fundamentals of NIR spectroscopy and NIR-HSI and its application in the global food and agriculture sectors are briefly considered. This review is concluded with an evaluation of bulk sample and single kernel analysis of wheat and triticale using NIR spectroscopy and NIR-HSI.

2.2 Wheat and triticale

Cultivation and selection of wheat strains started 10 000 years ago, as part of the 'Neolithic Revolution' (Shewry, 2009; 2018). The earliest cultivated form of grasses were diploid (genome AA) einkorn and the tetraploid (genome AABB) emmer wheat which originated in the south-eastern part of modern Turkey (Tanno and Willcox, 2006; Shewry, 2009). The major wheat species today is *Triticum aestivum L.*, an allohexaploid (2n = 6x =42) with three genomes A, B and D. Globally it accounts for more than 95% of the more than 700 mega-tonnes of wheat produced annually (Shewry and Hey, 2015).

Crossings between *Triticum turgidum* ssp. *durum* (AABB genome), durum wheat and *Aegilops tauschiii (DD genome)* resulted in *Triticum aestivum* (AABBDD genome) (Orth and Shellenberger, 1988; Monneveux *et al.*, 2000; El Baidouri *et al.*, 2017). The *Aegilops* species which is diploid and tetraploid in nature carries the U genome, enhancing abiotic and biotic stress resistance traits in bread wheat (Orth and Shellenberger 1988; Monneveux *et al*., 2000). The U genome therefore provides exploitable traits for plant breeders such as tolerance against drought, cold, heat and salt as well as elemental ion toxicity (Monneveux *et al*., 2000).

Bread wheat is further classified by its physical and chemical properties with common classification into hard and soft wheat varieties; where the terms hard and soft refer to the amount of force required to crush the wheat kernel. Hard wheat is used in bread and pasta production and soft wheat in biscuit making (Gazza *et al*., 2011; Quayson *et al*., 2016). On this basis, wheat differs in terms of physicochemical and functional properties, application, nutritional content and also ultimately in commercial value (Van der Merwe and Cloete, 2018). Bread wheat is also classified in terms of growing season, i.e. winter or spring

Durum wheat is known to be extremely drought tolerant, making it suitable for growing in Mediterranean areas with low annual rainfall. Furthermore, it is classified as a very hard wheat with a high protein content (Mohammadi 2016; Al Khateeb *et al*., 2017). This is due to the species not having the D genome (AABB) in contrast to bread wheat (AABBDD) (Quayson *et al.*, 2016).

Triticale is an intergeneric hybrid of wheat (*Triticum* spp.) and rye (*Secale cereale* L.) which was developed in 1875 by Stephen Wilson and perfected in 1888 by the German breeder A.D.W. Rimpau

(McGoverin *et al.*, 2011; Eudes, 2015; Zhu, 2018). Triticale was developed in order to combine the positive attributes of the parent species into a single plant, i.e. the breadmaking capabilities of wheat and rye being optimally suited for less premium growing conditions (McGoverin *et al.*, 2011). Triticale divergent varieties exhibit amphiploidy with respect to wheat (AABBDD) and rye (RR) genomes. However unalterable tetraploid, hexaploid (AABBR/D) and octoploid (AABBDDRR) triticale varieties have been bred. This is dependent on which parent is more pronounced in the crossing procedure. If bread wheat (AABBDD) is more prominent in the cross, octoploid triticale will be dominant. Hexaploid triticale is found when durum wheat (AABB) is crossed with rye (RR). In this manner specific traits can be selectively highlighted and can be taken advantage of by plant breeders (McGoverin *et al.*, 2011; Eudes, 2015; Cornejo-Ramírez *et al.*, 2016).

Triticale derived its drought tolerance from its parent species rye, making it suitable for growth in water sparse areas (Giunta *et al.*, 1993). Modern triticale varieties are on par with wheat varieties in terms of yield and in some cases triticale outperforms wheat when planted in marginal or barren soils (Mergoum *et al*., 2004). Usage of triticale hybrids are determined by its chemical composition (McGoverin *et al*., 2011). The composition of triticale being closer to wheat than rye, is reflective in the genome of triticale – two from wheat (A and B) and one from rye (R genome) (Varughese *et al*., 1996). The latter results in triticale not being ideal for breadmaking as the R or sticky gene derived from rye brings about poor breadmaking characteristics (McGoverin *et al.*, 2011).

2.2.1 Kernel morphology

Morphologically, wheat kernels appear oval, elliptical, elongated and truncated if viewed from the dorsal position. In North America, the average weight of the wheat caryopsis is 35 mg whereas European wheat weighs 55 mg on average. The outer dimensions are 2.0-3.0 mm (height) by 3.0-3.5 mm (width) by 6.0-8.0 mm (length) (Delcour and Hoseney, 1986; Gegas *et al*., 2010). The wheat caryopsis is rounded on the dorsal side and has a longitudinal fold running along the ventral side. This crease runs for nearly the entire length and extends to close to the centre of the kernel. The germ is located on the dorsal side, and oblique to this, hairs or the brush of the kernel is located. The colour of the outer pericarp is described as being either white or red. This is related to the anthocyanin content of the seed coat. These phenotypical properties enable identification of varieties.

The morphological structure of triticale follows that of its parent species closely (Góral *et al*., 2015). It has a crease on its ventral side and is rounded on its dorsal side. The overall length of the caryopsis is 10 to 12 mm with a width of 3 mm, giving an average weight of 40 mg per kernel. The caryopsis of triticale is, in general, longer than that of wheat, deriving its length from the rye parent. The colour of the grain is described as being yellow-brown, and the pericarp is characterised as having folds or waves caused by shrivelling. In some instances, if the parent rye species shows dominant signs of blue anthocyanin expression in the pericarp and purple in the aleurone layer, this expression can also be present in newly formed triticale kernels (Doshi *et al*., 2007; Li *et al*., 2011; Lachman *et al*., 2017).

2.2.2 Pericarp

The pericarp is composed of a multitude of functional layers, i.e. an outer epidermis, hypodermis, parenchyma, intermediate cells, cross cells and tube cells (Delcour and Hoseney, 1986). The intermediate and tube cells do not cover the kernel completely. Tube cells are long and cylindrical (125 × 20 µm) in dimension, and they are orientated with their long axis being perpendicular to the long axis of the caryopsis (Delcour and Hoseney, 1986). The seed coat and nucellar epidermis are joined to the tube cells on their distal and proximal sides, respectively. The three layers that make up the seed coat are a thick outer cuticle, a pigmented layer and a thin inner cuticle. The seed coat's thickness varies between 5 and 8 µm and the nucellar epidermis is about 7 µm thick (Delcour and Hoseney, 1986). The wheat pericarp makes up about 5% of the total kernel mass, comprising 20% cellulose, 6% protein, 2% ash and 0.5% fat. With the remainder of the pericarp consisting of non-starch branch-chained polysaccharides (Delcour and Hoseney, 1986)**.** The chemical makeup of the pericarp of triticale is similar to that of both its parents, wheat and rye (Wrigley *et al.*, 2016).

2.2.3 Endosperm

The aleurone layer encloses the starchy endosperm and thus forms the outermost layer of the endosperm. It is only one cell layer thick and the cells are distinct from starchy endosperm cells (Buttrose, 1963; Fulcher *et al*., 1972; Delcour and Hoseney, 1986). The aleurone cells are block shaped (37-65 µm × 25-75 µm) when viewed longitudinally, with thick cell walls (6-8 µm) that thin out as they move closer to and around the germ (Delcour and Hoseney, 1986).

During milling, the aleurone layer is removed as it is in direct contact with the bran (pericarp) layer. It has an abundance of chemical constituents which include high enzyme activity, ash, protein, total phosphorus, phytate phosphorus and lipid content (Delcour and Hoseney, 1986). In order to reduce the endosperm during milling into flour, farina (bread wheat) or semolina (durum wheat) the wheat variety and hardness of the kernels being milled have to be considered.

The cells present in the starchy endosperm are classified according to their geometrical conformation and their location. Sub-aleurone (peripheral) cells are those adjacent to the aleurone layer and they are similar in size (60 µm in diameter) to the aleurone cells (Khan and Shewry, 2009). Adjacent to the sub-aleurone layer, cells are made up and occupied by elongated prismatic starchy endosperm cells (150 × 50 × 50 µm) which extend inwards to the centre of the caryopsis crease. The centre of the starchy endosperm comprises generally round and polygonal starch cells, these are 72-144 µm in length and 69-120 µm in width (Delcour and Hoseney, 1986; Khan and Shewry, 2009). Endosperm cells that are closest to the aleurone layer are high in protein (up to 54%) while central cells are high in starch. The progressive starch gradient towards the centre of the endosperm causes dilution of also other components (minerals, vitamins, enzymes and various polyphenols) and not only protein (Delcour and Hoseney, 1986; Khan and Shewry, 2009)

Endosperm cell walls are mainly composed of 15% protein and 75% polysaccharide of which the latter comprises *ca*. 70% arabinoxylans, 20% (1→3,1→4)-β-D-glucan, 7% β-glucomannan and 2% cellulose (Bacic and Stone, 1980; Khan and Shewry, 2009). With cell wall size and composition depending on cell location within the endosperm relative to the exosperm.

Contained within the endosperm cells are starch granules embedded in a protein network; these form the two main energy storage reserves for the cell at maturity. The protein matrix is mostly, but not entirely, made up of glutenins and gliadins in their native form. These are found in a compressed form with mud- or clay-like appearance. Starch within the endosperm cells comprise large, lenticular granules of up to 40 µm across and spherical granules between 2 and 8 µm in diameter.

2.2.4 Embryo

The embryo or germ is positioned on the lower dorsal side of the caryopsis, perpendicular to the brush and comprises two major components, i.e. the embryonic axis and the scutellum. The scutellum forms the storage organelle and the embryonic axis the rudimentary root and shoot of the plant (Delcour and Hoseney, 1986; Khan and Shewry, 2009). A relatively high concentration of protein (25%), polysaccharides (18%), lipids (embryonic axis 16% and scutellum 32%) and ash (5%) are found in the wheat kernel embryo. There are no starch present in the embryo, but high levels of both water and fat soluble B and E vitamins.

2.3. Protein in small grains

Protein content has a significant impact on the final selling price of small grains, with many countries adopting it as a quality parameter in grading (Caporaso *et al*., 2018). Wheat proteins are of the most important components governing breadmaking with a protein content of up to 14% being ideal. In South Africa a protein content of between 11 and 12.5% is required whilst in Europe it ranges between 9 and 12%. Nevertheless, protein quantity alone cannot explain the differences in breadmaking quality (Weegels *et al*., 1996). Protein quality is of importance as it influences gluten formation during breadmaking.

In the early part of the 20th century the first report was given on the fractioning of cereal proteins (Osborne, 2011) – protein extraction of flour with a salt solution was done and two fractions were obtained, i.e. albumin (water soluble) and globulin (non-water soluble). Globulin was purified using dialysis and prolamins could be precipitated and extracted with an aqueous ethanol solution (70% v/v). Glutelin could also be extracted from the flour and salt solution using a dilute acetic acid solution. The extracted proteins (albumin, globulin, prolamin and glutelin) are referred to as the Osborne classes of protein (Osborne, 2011).

The functional proteins present in rye and triticale are similar to those found in wheat, however the functional proteins in rye and triticale do not form a viscoelastic dough. Triticale has a similar protein composition as its parent species rye. The water- and dilute-salt soluble proteins (albumin and globulin) are lower than for rye, whilst the prolamins are higher. In rye the albumins comprises *ca.* 35% and the globulins 10% of the total kernel protein. The prolamins constituted 20% and the acid soluble glutelins *ca.* 10% of the total protein. Around 20% of the total albumins and prolamins are is solubilised by the Osborne dilution scheme (Delcour and Hoseney, 1986).

2.4 Protein content determination in wheat

The Dumas combustion method detects total nitrogen content in an organic matrix. The sample is combusted at high temperature (950°C) in an oxygen rich atmosphere and through subsequent oxidation and reduction tubes the nitrogen is converted to N_2 gas. Secondary volatiles are trapped or separated through a series of scrubbers and nitrogen gas is finally measured by a thermal conductivity detector (Beljkaš *et al*., 2010). The results of which are given as percentage nitrogen or nitrogen as weight (mg) and this is then converted to protein percentage by using a conversion factor of 5.7 (in the case of wheat). The method allows for semi-automation and analyses time is shortened to five minutes per sample and it avoids the use of hazardous chemicals. This is compared to the Kjeldahl method which takes up to an hour or more to complete and uses concentrated sulphuric acid and a catalyst for acid digestion of samples. The Kjeldahl method determines only organic nitrogen and ammonia whilst the Dumas method determines total nitrogen including inorganic fractions such as nitrite and nitrate. Globally there is a clear trend to rather use the Dumas combustion method. Both of these methods have substantial running costs and they are destructive in nature, even if only a small sample (100 mg) is used. A more rapid, non-invasive and conclusive technique with a wider application is near-infrared spectroscopy (NIR) spectroscopy (Müller, 2017).

2.5. Endosperm texture and kernel hardness

Cereal endosperm texture is an important factor in small grains such as wheat as it determines its end use (Turnbull *et al.*, 2003). Wheat endosperm texture is genetically governed and described as either hard or soft. Environmental factors will only have an effect on the vitreousness and mealy (opaque or floury) appearance of the kernel. The degree of vitreousness is related to the packing density of the starchy endosperm. A tightly packed endosperm will be more vitreous than one which is loosely packed, resulting in a mealy visual appearance (Stenvert and Kingswood, 1977; Delcour and Hoseney, 1986). Hard wheat has a higher protein content than soft wheat, which in turn is rich in starch.

Kernel hardness is defined as the resistance to plastic strain and cracking with an applied force concentrated on the surface of the grain (Greenaway, 1969; Salmanowicz *et al.*, 2012). Various techniques are described to determine overall kernel hardness. These are divided into static and dynamic methods. Static methods include the measurement of the micro-hardness specific to cereals (Gasiorowski and Poliszko, 1977). Dynamic measurements of importance to the cereal industry include wheat hardness index (WHI) (Greenaway, 1969), particle size index (PSI) (Stenvert, 1974) and the pearling resistance index (PRI). Hardness can also be determined using NIR spectroscopy with the added advantage of being non-invasive (whole kernels), rapid and specific if milled wheat is used (Delwiche, 1993; Manley *et al.*, 2002a; Maghirang & Dowell, 2003; Dagou & Richard, 2016). Another common method used for kernel hardness analysis is the Single Kernel Characterisation System (SKCS) (Gaines *et al.*, 1996; Osborne and Anderssen, 2003; Muhamad and Campbell, 2004; Edwards *et al.*, 2007).

The earliest work on defining and recognising the difference in texture among grain lots dates back to the late 1800s (Cobb, 1896). In the second half of the 20th century work started on the commercial viability of cereal grains. This highlighted the need for genetic studies to be conducted for the mode of texture inheritance in cereals. Early work in the mid-1970s revealed that the major contributor to kernel texture was the effect of a single gene on grain texture (Mattern *et al.*, 1973; J. and Dyck, 1975). The genetic basis of endosperm hardness focusses on the Hardness (Ha) locus, which is located on chromosome 5D. It was further designated that the soft allele would be Ha and the hard allele ha (Mattern *et al.*, 1973).

In 1986, Philip Greenwell and J.D Schofield spearheaded a new notion when they extracted a Mr 15 kDA protein from water-washed wheat flour, and molecularly separated the proteins by gradient SDS-PAGE (Greenwell and Schofield 1986). They showed that the presence of the 15 kDa protein was associated with soft wheat and bound to the endosperm starch. This was confirmed in more than 150 different wheat varieties, including seven durum varieties. Greenwell and Schofield (1986) showed a linear relationship between the adhesion strength and the concentration of this specific protein in the endosperm. This protein was subsequently named 'friabilin', highlighting the fact that soft wheats are more friable than hard wheat (Greenwell and Schofield, 1989).

In the 1990s, evidence was found that friabilin is not made up of a single protein, but rather that it consists of multiple polypeptides (Jolly *et al.*, 1993; Morris *et al.*, 1994; Oda, 1994). It was suggested that some friabilin polypeptides may be puroindoline polypeptides (Jolly *et al.*, 1993), i.e. puroindoline a (Pin-a) and b (Pin-b) (Salmanowicz *et al.*, 2012). Grains that are soft have more of the wild allele gene encoding for Pin-a, and they accumulate both of the puroindoline on the surface of starch granules. Mutated alleles at Pin-b are found in medium and hard wheats. This results in a reduced amount of Pin-b on the starch granules (Salmanowicz *et al.*, 2012).

The milling industry regards the endosperm texture of small grains as important, as it directly correlates to milling quality, flour yield and financial gain. Hard grains results in a course flour with high amounts of damaged starch, whilst soft grain produces a fine flour with a lower degree of starch damage (Bolling, 1987). This is due to the point fracture within the endosperm – in hard grain the starch granules are cleaved and in soft grain the fractioning takes place between the starch granules.

2.5.1 Kernel hardness determination methods

Wheat hardness measurements go as far back as 1896 when a pair of pinchers was used to cut a wheat kernel in half, simulating the biting force of vertical and lateral incisors (Cobb, 1896). During the mid-1980s it became important to measure the difference between soft and hard wheat species. It became more difficult to visually inspect for hardness differences, as the crossing of cereal lines became ever more advanced (Miller *et al*., 1981; Sampson *et al*., 1983; Pomeranz *et al*., 1984; Lai *et al*., 1985; Gaines, 1986; Mattern, 1988).

Principles used to analyse, predict and measure kernel hardness are based on fractioning resistance (SKCS) (Gaines *et al.*, 1996), sieving by particle size index (PSI) (Symes, 1965) and scattering of NIR radiation on the whole kernel (Maghirang and Dowell, 2003) and flour (Osborne *et al.*, 1981; Manley *et al.*, 2002b; Armstrong *et al.*, 2006). The Pohl Farinator or hardness cutter is also commonly used.

2.5.2 Single Kernel Characterisation System (SKCS)

The Single Kernel Characterisation System (SKCS) model 4100 (Perten Instruments, North America, Inc., Reno, NV) is used and designed for the classification of wheat into four ranges based on the hardness or softness of the kernel (Martin *et al.*, 1993; Gaines *et al.*, 1996). The SKCS instrument is designed to isolate individual kernels (ca. 300, 15 g), weigh them and then crush them between a rotor and crescent gap. Conductivity between the motor and the crescent-shaped gap is measured and also the deformation profile of the kernel. This information is then mathematically calculated to provide the average weight, size, moisture content and hardness of the sample.

Processing of 300 kernels takes *ca.* three minutes – the method can thus be classified as a rapid technique (Gaines *et al.*, 1996). Results obtained are given in terms of hardness index (HI) which relates to hard wheat requiring greater force to be crushed than soft wheat. In Table 2.1 the average HI values is given for different hardness categories (Gaines *et al.*, 1996; AACC Approved Methods of Analysis, 1999a).

Table 2.1 Hardness index categories for soft to hard kernels as adapted from AACC International method 55-31.01 (AACC Approved Methods of Analysis, 1999a)

Hardness Category	**HI**[a]	**PSI**[b]
Extra Soft	0-10	76+
Very Soft	10-24	71-75
Soft	25-34	67-70
Medium Soft	35-44	63-66
Medium Hard	45-64	58-62
Hard	65-80	50-57
Very Hard	81-90	40-50
Extra Hard	91 +	35-40

[a] Hardness index [b] Particle size index, cyclone ground kernels

2.5.3 Pohl Farinator

The Pohl Farinator test is used to determine hardness of kernels based on their vitreousness. The method is according to International Association for Cereal Science and Technology (ICC) standard method 129 (Anon, 1980). This involves 100 random whole kernels being sampled from a consignment or batch, and subsequently cut in half where their vitreousness or non-vitreousness is assessed visually. Vitreousness is calculated as follows and described in detail by Branković *et al.*, 2014.

$$\text{Grain vitreousness (\%)} = A + \tfrac{3}{4} B + \tfrac{1}{2} C + \tfrac{1}{4} D$$

Where,

A = number of fully vitreous grains

B = number of vitreous grains with more than 75% of grain cross-section being vitreous

C = number of vitreous grains with 50% to 75% grain cross-section being vitreous

D = number of vitreous grains with 25% to 50% grains cross-section being vitreous

The Pohl Farinator test is found to be imprecise due to subjective operator behaviour and due to the nature of binomial data (Wesley *et al.*, 2005). Not only is the technique biased and statistically uncertain, it is also a destructive method which will not be suitable to be used in breeding programmes when only a small amount of sample is available.

2.5.4 Particle size index (PSI)

The particle size index (PSI) (Symes, 1965; Stenvert, 1974) test is described by AACC method 55-30.01 (AACC Approved Methods of Analysis, 1999b). It is based on determining the relative hardness of a small grain sample by grinding and sieving. A hard small grain will produce a flour with large particle sizes and a lower percentage throughs, resulting in a lower PSI value. The method involves weighing the flour that has moved through the sieve. The PSI is then expressed as the percentage throughs. In Table 2.2 the average PSI values are shown for wheat ranging from extra soft to extra hard. The PSI method is not a rapid method and is not suited for industry application. It is, however, a very precise method and is used as a reference method and for calibration of other methods, e.g. NIR spectroscopy.

Table 2.2 Average particle size index (PSI) values for different hardness categories of wheat. (AACC Approved Methods of Analysis, 1999b)

Hardness Category	PSI (%)
Extra Soft	> 35
Very Soft	31-35
Soft	26-30
Medium Soft	21-25
Medium Hard	17-20
Hard	13-16
Very Hard	8-12
Extra Hard	0-7

2.5.5 Near-infrared spectroscopy – kernel hardness

NIR reflectance spectroscopy provides for a rapid, non-invasive method for compositional factors in ground samples of grain. In accordance with the AACC method 39-70.02 (AACC Approved Methods of Analysis, 1999c) it is advised to use reflectance spectroscopy on a ground grain sample. NIR reflectance signal is affected by particle size distribution of ground grain, with NIR absorption increasing with grain hardness (larger particles). The difference in flour particle size influences the amount of NIR radiation scattered within the sample. The large particles absorb more incident radiation than smaller particles, thus it has a higher energy absorbance value (Pomeranz and Williams, 1990).

Using Fourier transform NIR (FT-NIR) spectroscopy, kernel hardness has been predicted on whole wheat flour (Manley *et al*., 2002b). Whole kernel hardness using NIR spectroscopy has also been done (Williams, 1991; Dowell, 2000; Maghirang and Dowell, 2003).

2.6 Near-infrared spectroscopy

Frederick William Herschel discovered the first non-visible region in the electromagnetic absorption spectrum, i.e. NIR (Herschel, 1832). This region was, however, not considered to be of analytical importance for another 150 years. In the interim, scientific focus and methods used revolved around conventional techniques, such as gravimetrical analysis – oven drying for moisture analysis and Kjeldahl for protein determination. Since 1949, the NIR technique was revived by Karl Norris and subsequently Phil Williams applied it into a practical method and showed the potential of this rapid technique being applied to small grains (Norris, 1996; Williams *et al*., 2019). Scientific work done with NIR technology, through the period 1800 to 2003, has been extensively reviewed (McClure, 2003). More recently, the application of NIR spectroscopy and hyperspectral imaging for the analysis of biological materials (Manley, 2014), authentication of foods (Manley & Batten, 2018; Wang *et al.*, 2017), food safety evaluation and control (Qu *et al.*, 2015) and the quality and safety evaluation of cereals (Sendin *et al*., 2018) has been shown.

Being a secondary method, NIR spectroscopy requires reference values for calibration and validation. Thus NIR methods depend on the accuracy and precision of reference methods such as Kjeldahl or Dumas combustion for protein and air oven methods for moisture content determination. In contrast to these methods, NIR technology is non-invasive, rapid, chemical free and easy to use, provided that an established method and model has been developed and proven to be robust (Manley 2014)

Near-infrared hyperspectral imaging (NIR-HSI) is not a new concept. The term was first used by Goetz *et al*. in 1985 for remote sensing applications (ElMasry *et al*., 2012). It was only during the late 1990s that this technology became available to the academic research sector and public domain for food and agricultural applications. The advantage of NIR-HSI is that it combines NIR spectroscopy with digital imaging – this enables both spatial and spectral data to be obtained simultaneously (Gowen *et al.*, 2007). In conventional NIR spectroscopy only an average spectrum is obtained from the sample scanned.

2.6.1 Fundamental principles of near-infrared spectroscopy

Near-infrared spectra result from the energy absorption and subsequent vibration of molecular bonds in organic molecules. These comprise of overtones and combinations of overtones originating from vibrations occurring in the mid-infrared (MIR) region of the electromagnetic spectrum (Kirchler *et al.*, 2017). The MIR region is of higher energy than the NIR region, making for a decrease in signal intensity for NIR spectra (Manley, 2014).

The NIR region extends from 780 to 2500 nm (12500 to 4000 cm^{-1}) falling between the visible 380 to 780 nm (26316 to 12820 cm^{-1}) and MIR 2500 to 15000 nm (4000 to 400 cm^{-1}) regions (Porep *et al*., 2015). The main energy absorbers in the NIR region involve the energy response of chemical bonds such as O-H, C-H, C-O and N-H. This vibrational energy change is translated into an absorption spectrum within the NIR spectrophotometer (Cen *et al*., 2016)

Three common sensing modes for spectral analysis exist, namely reflectance, transmittance and interactance (Fig. 2.1). In reflectance the detector captures light reflected from the illuminated sample with a specific angle as to avoid specular reflection. In transmittance mode the detector and light sources are located on opposite sides of the sample being scanned or imaged. The detector captures the light which has been transmitted through the sample and is generally acquired as absorbance values. This method carries more valuable internal information, it is however dependent on sample thickness, density and composition. Transmittance mode is used to detect internal component concentration and to detect relevant characteristics of transparent materials. On the basis of a transmittance setup, the interactance mode can detect more information from the sample and is less hindered by surface scattering effects compared to reflectance as the light source is indirect to the imaged object by means of a light seal. Interactance mode also reduces the influence of sample thickness which offers a practical advantage over transmittance mode, however, it is limited when application involves high conveyor speed (ElMasry and Sun, 2010; Wu and Sun, 2013; ElMasry and Nakauchi, 2016).

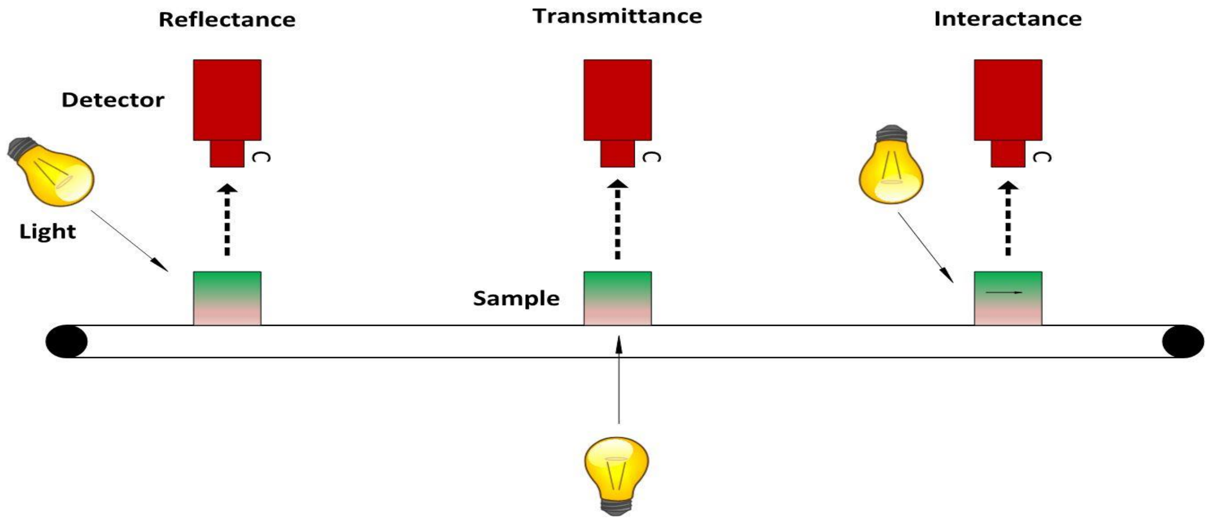

Figure 2.1 NIR spectroscopy and classic NIR-HSI sensing modes including reflectance, transmittance and interactance modes of detection.

2.6.2 Chemometrics

Due to the low absorbance frequency, overtone and combination modes, high interactance and overlap of possible chemical vibrations and high instance of spectral noise, NIR spectra are complex to interpret. This complexity are mainly due to overlapping, broad bands (multicollinearity). An indirect approach for extracting attainable data from the spectra are thus required, as visual inspection does not offer enough information about specific chemical features and information hidden within the spectra. This hurdle is overcome through the use of appropriate regression techniques which determines relationships between absorption values at specific wavelengths and quantitative reference values. The proposal to use multiple linear regression (MLR) to analyse NIR spectral data was made by Norris in the late 1960's (Norris, 1996), later being aptly termed chemometrics.

Exploratory data analysis is often performed using principal component analysis (PCA) (Rinnan *et al.*, 2009; Rinnan, 2014).Regression techniques most often used in NIR data analysis are principal component (PCR) and partial least squares (PLS) regression. Common classification techniques include partial least squares discriminant analysis (PLS-DA), linear discriminant analysis (LDA), factor analysis (FA) and cluster analysis (CA).

2.7 Fundamentals of hyperspectral imaging

Non-invasive imaging techniques such as hyperspectral imaging and red green blue (RGB) imaging are extremely advantageous for online, at-line or inline inspection of food and other agricultural commodities. Table 2.3 compares the differences between conventional NIR spectroscopy, hyperspectral imaging, RGB imaging and multispectral imaging. The practical advantages of hyperspectral imaging are highlighted in the table and it is shown that it exceeds at being a flexible multidimensional method.

Table 2.3 Differences between conventional NIR spectroscopy, hyperspectral imaging, RGB imaging and multispectral imaging (adapted from Wu and Sun, 2013**)**

Features	Conventional NIR Spectroscopy	Hyperspectral imaging	RGB Imaging	Multispectral imaging
Spectral information	√	√	X	Limited
Spatial information	X	√	√	√
Multi-constituent information	√	√	X	Limited
Detectability to objects with small size	X	√	√	√
Flexibility of spectral extraction	X	√	X	√
Generation of quality attribute distribution	X	√	X	Limited

2.7.1 Hyperspectral image acquisition

Hyperspectral images can be acquired in four different ways: line-by-line spatial scanning (pushbroom imaging); point-to-point spectral scanning (whisk-broom imaging); area scanning (staring imaging, tuneable filter or wavelength scanning); and also the single shot method (Wu and Sun, 2013; ElMasry and Nakauchi, 2016). Pushbroom image acquisition involves a whole image line and spectral information corresponding to spatial pixel position to be obtained. Due to the scanning of an object through the spectral lines, this type of image acquisition is suitable for conveyor belt systems that are commonly used in food production. For the whiskbroom technique, a single point (pixel) is scanned at a time, providing the spectrum at this point. Subsequent points are scanned by moving the object or the detector along the spatial direction coaxially to the detector or object, depending on which is being moved. Area scanning is a spectral scanning method, which keeps the image field of view fixed and acquires a 2-D image with x and y directions. Giving full spatial information at a single wavelength at a time, resulting in a stack of single band images. This technique is suitable for applications where the object can be stationary for a period of time. The single shot method gathers both spectral and spatial information with a large area detector with one exposure to capture the spectral images. This is an attractive solution when rapid hyperspectral imaging is required for a multispectral instrument (Wu and Sun, 2013).

Hyperspectral images are known as hypercubes, which are built up from hundreds of single channel grayscale images, each layer consisting of pixels and spectral data. Hypercubes are three dimensional superimposed data matrices, consisting of two dimensional images composed of pixels in the x and y direction and wavelength dimension in the z direction (ElMasry and Sun, 2010; Wu and Sun, 2013; Qu *et al.*, 2015; Liu *et al.*, 2017; Munir *et al.*, 2018).

Imaging equipment for hyperspectral imaging is costly, especially when wavelengths of up to 2500 nm are required. The wavelengths between 1100 to 2500 nm require the more costly indium gallium arsenide (InGaAs) or mercury cadmium telluride (HgCdTe) based array detectors. For wavelengths up to 1100 nm, which include the visible light range, silicon based detectors which are lower in cost, can be used (Manley 2014).

Image analysis is performed after a data cube is obtained, subsequently dead pixels, spectral spikes and background is removed to allow only for the region of interest. Exploratory spectral pre-treatment and compression by PCA is performed to obtain a corrected data cube on which further chemometric techniques are applied or spectral data can be extracted. A tutorial for hyperspectral image analysis has been published by Amigo *et al.*, 2015 and it showcases the practical aspects behind spectral imaging.

2.8 Near-infrared spectroscopy and hyperspectral imaging of small grains

Application of NIR spectroscopy to quantitatively predict chemical and physical attributes of small grains such as moisture (Hruschka and Norris, 1982; Windham *et al.*, 1997), protein (Orman and Schumann, 1991; Kays *et al.*, 2000; Jimenez *et al.*, 2019), lipids (Chen *et al.*, 1997; Vines *et al.*, 2005; Wang *et al.*, 2006; Saleh *et al.*, 2008) and hardness (Manley *et al.*, 2002a, 2011; Mahesh *et al.*, 2014; Caporaso *et al.*, 2018; Ibrahim *et al.*, 2018) has been demonstrated.

2.8.1 Conventional NIR spectroscopy

Initially, spectroscopic data were collected from ground cereals (Williams *et al.*, 1978). Grinding provides for a more uniform material and predictions using conventional NIR spectroscopy are usually better, compared to whole kernel measurements. Grinding, however, is tedious and presents a time limitation when a large sample set needs to be scanned. In addition, milling of grains removes information related to the natural chemical fluctuations of individual kernels within a batch, as it implies providing an average result (Caporaso *et al.*, 2018). It is now recognised that reliable predictions of whole wheat kernel composition is possible with NIR spectroscopy with the advantage of no sample preparation (Williams, 1991; Williams and Sobering, 1993).

The suitability of NIR spectroscopy for the analysis of single kernels was shown (Delwiche, 1995, 1998; Fox and Manley, 2014). Single wheat kernel analysis using NIR transmittance was done to determine whole kernel protein content (Delwiche, 1995). In this study six wheat classes, i.e. hard red winter, hard red spring, hard white, soft red winter, soft white and durum were examined. Of these, five samples per class were taken, with each sample comprising 96 randomly selected wheat kernels. The average single kernel spectra (850-1050 nm) was subsequently used to develop partial least squares (PLS) regression models for protein content prediction. The reference data was collected on each kernel by means of the Dumas combustion technique. Model accuracies (R^2) ranged between 0.85 and 0.93 and standard errors of prediction (SEP) between 0.4 and 0.9% (Delwiche, 1995).

Classification of five wheat classes using PLS and multiple linear regression for single kernels has also been focussed on (Delwiche and Massie, 1996). Single kernel NIR reflectance scans with two spectral regions (551-750 for colour and 1120-2476 nm for intrinsic property distinctions) were taken on 10 randomly drawn kernels from 318 commercially sourced samples. Classification was done on five wheat classes, i.e. hard white, hard red spring, hard red winter, soft red winter and soft white through PLS and multiple linear regression (MLR) analyses to develop binary decision models. With a five-class model prediction accuracy being greatest when red wheat and white wheat varieties were compared, indicating that wheat colour was dominating the classification (Delwiche & Massie, 1996).

In the early 2000's it was shown that bulk sample protein could be predicted from single kernel NIR spectral readings (Delwiche & Hruschka, 2000). Five wheat classes were used from which 10 kernels were randomly selected to make up a total sample size of 318 single kernels. The study showed that with as few as 300 wheat kernels, bulk protein content from single kernel (SK) spectra could be accurately predicted – equivalent to that of conventional bulk NIR instrumentation.

The same year saw research focussing on differentiating between vitreous and non-vitreous durum wheat kernels by using NIR spectroscopy. With classification accuracy being 72% for the prediction set and 73% for the calibration set from a sample set of 240 single kernels. From the 240 kernels, 80 kernels were selected which were determined to be ‘obvious vitreous or non-vitreous’ and a 100% prediction and classification accuracy was achieved (Dowell 2000).

The development of a non-invasive method for protein content, vitreousness, density and hardness index for single kernels of European wheat was also done (Nielsen *et al*., 2003). Using NIR spectroscopy in transmission mode a less than adequate calibration for hardness index was obtained (R^2 0.59, RMSEP 20.2). For protein content the prediction results were more in line with other studies and a R^2 of 0.98 and RMSEP of 0.48 was obtained. The needs of wheat breeders were realised when a non-destructive NIR method was developed to segregate single wheat kernels based on a high and low protein values. This was achieved by equipping a commercial colour sorter with NIR filters. With results showing that sorting was mainly driven by colour and vitreousness of the wheat kernels (Pasikatan & Dowell, 2004).

NIR spectroscopy work on triticale is limited. Igne *et al.* (2007) created a prediction model for protein and moisture content of bulk triticale grain. They determined that existing wheat models were not applicable for moisture content prediction with SEP_{avg} = 0.37% for triticale compared to 0.15% for wheat. However, existing wheat models were more applicable for screening of protein content with SEP_{avg} = 0.38% for triticale compared to 0.25% for wheat. To achieve better prediction results, dedicated triticale calibrations were developed, this gave better prediction results than using wheat calibrations (Moisture: SEP 0.19-0.50%, Protein: SEP 0.22-0.68%) for triticale predictions (Moisture: 0.15-0.29, Protein: 0.30-0.34). The authors had a large sample set of 412 for moisture and 502 for protein content which was highly suitable for calibration of a robust model, however their recommendation was still to use individual dedicated models for the determination of triticale moisture and protein content. The authors also concluded that it would be suitable for the triticale spectral data set to be added to the wheat prediction models to obtain better prediction accuracy.

Manley *et al.* (2013) predicted triticale grain quality parameters based on both chemical and indirectly measured reference methods, using NIR spectroscopy. NIR spectroscopy calibrations for determining protein, moisture and ash contents as well as kernel hardness were performed. Prediction models were best for milled samples compared to whole grain samples. The best calibration results were obtained on direct chemical reference measurements (protein and moisture content), compared to those based on indirect measurements (PSI, ash content and SDS sedimentation). It was, however, stated that calibrations on indirect measurement were still useful to identify extreme samples which did not entirely fall within the model parameters. For ground grain a SEP of 0.52% (w/w) and coefficient of determination (R^2) of 0.95 was obtained while for whole grain prediction accuracies were less accurate with an SEP of 0.67% and R^2 of 0.92 (Manley *et al*., 2013).

2.8.2 NIR hyperspectral imaging

Detection of insect-damaged wheat kernels was evaluated (Singh *et al*., 2009). Wheat kernels were imaged in the 1000-1600 nm wavelength range using an NIR hyperspectral imaging system. The obtained images at 1101.69 and 1305.05 nm were subjected to statistical discriminant classifiers, i.e. linear, quadratic and Mahalanobis. Linear discriminant analysis and quadratic discriminant analysis were the most accurate and correctly classified 85 to 100% healthy and insect-damaged wheat kernels.

The diffusion of water through single wheat kernels of different hardness with regards to time was mapped using NIR-HSI (Manley *et al*., 2011). Contaminants such as foreign materials (barley, canola, maize, flaxseed, stones) were identified in Canada Western Red Spring wheat using NIR hyperspectral imaging. The classification model was developed using standard normal variate (SNV) as the pre-processing technique and k-nearest neighbours (k-NN) as the classifier. The calibration and validation error of the models were found to be similar with classification error being above 97% for all classes (Ravikanth *et al*., 2016).

Two regression techniques were compared by Mahesh *et al*. 2014 (PLSR and principal component regression (PCR)) for both protein content and hardness prediction using NIR-HSI of bulk wheat samples. With model results for PLSR modelling being 1.76, 1.33 and 0.68 for MSEP, SECV and R^2 and for the PCR model, 2.02, 1.42 and 0.62. For kernel hardness prediction using PLSR the values were 16.2, 4.03 and 0.88 for MSEP, SECV and R^2 and for the underperforming PCR model it was 22.6, 4.75 and 0.72. It was noted by the author that PLSR models significantly outperformed the PCR models. Better results could possibly be obtained by using PLSR as PCR only explains variability in the predicted variables by creating components without taking the response variable into account to lower the number of model components. PLSR takes the response variable into account to lower model complexity which often fits the response variable better (Næs and Martens, 1988; Wold *et al.*, 2001). Mahesh could also have expected better prediction accuracies if the wheat samples were imaged as is and not conditioned to different moisture levels and artificially increasing the sample size in this manner. As an adjustment in moisture levels is not specifically an adjustment towards the chemical nature of the wheat kernels, thus it can be concluded that the actual sample size was much smaller than what the author stated

Quantification of protein content in milled wheat has been shown, where NIR hyperspectral imaging was compared to conventional NIR spectroscopy (Morales-Sillero *et al.*, 2018). PLS calibration models were set up over the whole wavelength range for individual instruments and specifically for the common range (1120 -2424 nm). The models were validated using the leave-one-out cross validation procedure and it was validated using an independent validation set. Results showed that both instruments performed equally well when the common wavelength range was used. Giving an R^2-value of 0.99 for three instruments and root mean square error in prediction (RMSEP) values of 0.15% for NIR-HSI and NIR System DS2500 and 0.16% for the Perten instrument. This showed that there was no difference between the techniques used.

Protein prediction on single whole wheat kernels was performed, wheat samples from 2013 and 2014 harvest seasons were sourced from United Kingdom (UK) millers (Caporaso *et al.*, 2018). The samples were analysed by Dumas combustion and subsequently an NIR-HSI method for total protein content prediction was set up. The spectral region selected for HSI was 980-2500 nm in reflectance mode, using the pushbroom approach. Spectral data of single kernels were then used to develop partial least squares (PLS) regression models for protein content prediction of single kernels. Overall performance of the calibration model was evaluated using the R^2-value and root mean square error (RMSE) from 3250 calibration set and 868 validation set samples. This gave R^2-values of 0.82 and 0.79, and RMSE of 0.86 and 0.94 for the calibration and validation set, respectively. This enabled quantification of the protein distribution between single kernels, and pixel wise visualisation of the protein distribution within the kernels. The SK wheat protein content range of 6.2-19.8% used by Caporaso et al. (2018) shows that the lower and higher regions are underrepresented. Caporaso et al. (2018) could have achieved better calibration results using less LV's if an advanced spectral pre-treatment method such as GLS was applied to the authors data set.

2.9 Conclusion

The review shows the importance of understand the fundamental biochemical properties of wheat and triticale kernels. And it highlights the shortfalls of conventional analytical techniques that are used daily in the grain industry, which are reliable and will continue to be so. The industry is, however, tied up under the paradigm of outdated techniques which are expensive and time consuming. Conventional NIR spectroscopy and NIR hyperspectral image analysis of single and bulk cereal grain kernels have been shown to be a proven analytical technique to accurately, within model constraints, predict chemical properties quantitatively and qualitatively. It has been used to accurately and routinely predict protein content, moisture content and hardness attributes of wheat kernels. In addition, it has been applied to distinguish between wheat of different classes and to accurately distinguish between contaminants. NIR-HSI and conventional NIR-spectroscopy studies on triticale are limited. With no NIR-HSI work being performed on triticale, opening the field for work to be carried out. The increase of technological capacity also identifies the need for building prediction models that are more suited to the latest advancements in the field. The need for NIR-HSI models exists which quantitatively predict protein and moisture content and also kernel hardness of wheat and triticale whole grain, both for a bulk sample approach and on a SK level. The non-invasive nature of such a technique will also allow for the opening up of further more in detail approaches to NIR-HSI of whole grains and the application of the technique in industry.

2.10 References

AACC Approved Methods of Analysis. (1999a). *Method 55-31.01. Single kernel characterization system (SKCS) for wheat kernel texture*. 11th edn. St. Paul, MN, U.S.A.: Cereals & Grains Association.

AACC Approved Methods of Analysis. (1999b). *Method 55-30.01 Particle size index for wheat hardness*. 11th edn. St. Paul, MN, U.S.A.: Cereals & Grains Association.

AACC Approved Methods of Analysis. (1999c). *Method 39-70.02, Near-infrared reflectance method for hardness determination in wheat*. 11th edn. St. Paul, MN, U.S.A.: Cereals & Grains Association.

Amigo, J.M., Babamoradi, H. & Elcoroaristizabal, S. (2015). Hyperspectral image analysis. a tutorial. *Analytica Chimica Acta*, **896**, 34-51

Anon. (1980). ICC Standard Method No. 129 Method for the determination of the vitreousness of durum wheat.

Armstrong, P.R., Maghirang, E.B., Xie, F. & Dowell, F.E. (2006). *Comparison of dispersive and fourier-transform NIR instruments for measuring grain and flour attributes*. *Applied Engineering in Agriculture*, **22**, 453-457

Bacic, A. & Stone, B. (1980). A (1→3)- and (1→4)-linked β-d-glucan in the endosperm cell-walls of wheat. *Carbohydrate Research*, **82**, 372–377.

Baidouri, M. El, Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., Alaux, M., Quesneville, H., Pont, C. & Salse, J. (2017). Reconciling the evolutionary origin of bread wheat (Triticum aestivum). *New Phytologist*, **213**, 1477–1486.

Beljkaš, B., Matić, J., Milovanović, I., Jovanov, P., Mišan, A. & Šarić, L. (2010). Rapid method for determination of protein content in cereals and oilseeds: validation, measurement uncertainty and comparison with the Kjeldahl method. *Accreditation and Quality Assurance*, **15**, 555–561.

Bolling, H. (1987). Milling quality of wheat. *European Conference on Food Science and Technology*, 259–284.

Branković, G.R., Dodig, D., Zorić, M.Z., Šurlan-Momirović, G.G., Dragičević, V. & Durić, N. (2014). Effects of climatic factors on grain vitreousness stability and heritability in durum wheat. *Turkish Journal of Agriculture and Forestry*, **38**, 429–440.

Buttrose, M. (1963). Ultrastructure of the developing wheat endosperm. *Australian Journal of Biological Sciences*, **16**, 305.

Caporaso, N., Whitworth, M.B. & Fisk, I.D. (2018a). Protein content prediction in single wheat kernels using hyperspectral imaging. *Food Chemistry*, **240**, 32–42.

Caporaso, N., Whitworth, M.B. & Fisk, I.D. (2018b). Near-Infrared spectroscopy and hyperspectral imaging for non-destructive quality assessment of cereal grains. *Applied Spectroscopy Reviews*, **53**, 667–687.

Cen, H., Lu, R., Zhu, Q. & Mendoza, F. (2016). Nondestructive detection of chilling injury in cucumber fruit using hyperspectral imaging with feature selection and supervised classification. *Postharvest Biology and Technology*, **111**, 352–361.

Chen, H., Marks, B.P. & Siebenmorgen, T.J. (1997). Quantifying surface lipid content of milled rice via visible/near-infrared spectroscopy. *Cereal Chemistry*.

Cobb, N.A. (1896). The hardness of the grain in the principal varieties of wheat. *Agric. Gazette*, 279–298.

Cornejo-Ramírez, Y.I., Ramírez-Reyes, F., Cinco-Moroyoqui, F.J., Rosas-Burgos, E.C., Martínez-Cruz, O., Carvajal-Millán, E., Cárdenas-López, J.L., Torres-Chavez, P.I., Osuna-Amarillas, P.S., Borboa-Flores, J. & Wong-Corral, F.J. (2016). Starch debranching enzyme activity and its effects on some starch physicochemical characteristics in developing substituted and complete triticales (X Triticosecale Wittmack). *Cereal Chemistry*, **93**, 64–70.

Dagou, S. & Richard, F.C. (2016). Inheritance of kernel hardness in spring wheat as measured by near-infrared reflectance spectroscopy. *Euphytica*, **209**, 679–688.

Delcour, J.A. & Hoseney, R. (1986). *Principles of Cereal Science*. Third Edit. AACC International, Inc.

Delwiche, S.R. (1995). Single wheat kernel analysis by near-infrared transmittance: protein content. *Cereal Chemistry*, **72**, 11-16.

Delwiche, S.R. (1998). Protein content of single kernels of wheat by near-infrared reflectance spectroscopy. *Journal of Cereal Science*, **27**, 241-254.

Delwiche, S.R. & Hruschka, W.R. (2000). Protein content of bulk wheat from near-infrared reflectance of individual kernels. *Cereal Chemistry*, **77**, 86-88

Delwiche, S.R. & Massie, D.R. (1996). Classification of wheat by visible and near-infrared reflectance from single kernels. *Cereal Chemistry*, **73**, 399-405.

Doshi, K.M., Eudes, F., Laroche, A. & Gaudet, D. (2007). Anthocyanin expression in marker free transgenic wheat and triticale embryos. *In Vitro Cellular and Developmental Biology - Plant*, **43**, 429–435.

Dowell, F.E. (2000). Differentiating vitreous and nonvitreous durum wheat kernels by using near-infrared spectroscopy. *Cereal Chemistry*, **77**, 155–158.

Edwards, M.A., Osborne, B.G. & Henry, R.J. (2007). Investigation of the effect of conditioning on the fracture of hard and soft wheat grain by the single-kernel characterization system: a comparison with roller milling. *Journal of Cereal Science*, **46**, 64–74.

ElMasry, G., Kamruzzaman, M., Sun, D.W. & Allen, P. (2012). Principles and applications of hyperspectral imaging in quality evaluation of agro-food products: a review. *Critical Reviews in Food Science and Nutrition*, **52**, 999–1023.

ElMasry, G. & Sun, D.W. (2010). Principles of hyperspectral imaging technology. *Hyperspectral Imaging for Food Quality Analysis and Control*, 3–43.

ElMasry, G.M. & Nakauchi, S. (2016). Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality - A comprehensive review. *Biosystems Engineering*, **142**, 53–82.

Fox, G. & Manley, M. (2014). Applications of single kernel conventional and hyperspectral imaging near infrared spectroscopy in cereals. *Journal of the Science of Food and Agriculture*, **94**, 174–179.

Fulcher, R.G., O'Brien, T.P. & Lee, J.W. (1972). Studies on the aleurone layer: I. Conventional and fluorescence microscopy of the cell wall with emphasis on phenol-carbohydrate complexes in wheat. *Australian Journal of Biological Sciences*, **25**, 23–34.

Gaines, C.S. (1986). Texture (hardness and softness) variation among individual soft and hard wheat kernels. *Cereal Chemistry*, **63**, 479-484.

Gaines, C.S., Finney, P.F., Fleege, L.M. & Andrews, L.C. (1996). Predicting a hardness measurement using the single-kernel characterization system. *Cereal Chemistry*, **73**, 278–283.

Gasiorowski, H. & Poliszko, S. (1977). A wheat endosperm microhardness index. *Acta Alim*, **6**, 113–117.

Gazza, L., Sgrulletta, D., Cammerata, A., Gazzelloni, G., Perenzin, M. & Pogna, N.E. (2011). Pastamaking and breadmaking quality of soft-textured durum wheat lines. *Journal of Cereal Science*, **54**, 481–487.

Gegas, V.C., Nazari, A., Griffiths, S., Simmonds, J., Fish, L., Orford, S., Sayers, L., Doonan, J.H. & Snape, J.W. (2010). A genetic framework for grain size and shape variation in wheat. *Plant Cell*, **22**, 1046–1056.

Giunta, F., Motzo, R. & Deidda, M. (1993). Effect of drought on yield and yield components of durum wheat and triticale in a Mediterranean environment. *Field Crops Research*, **33**, 399–409.

Goetz, A.F.H., Vane, G., Solomon, J.E. & Rock, B.N. (1985). Imaging spectrometry for earth remote sensing. *Science*, **228**, 1147–1153.

Góral, H., Stojałowski, S., Warzecha, T. & Larsen, J. (2015). *The development of hybrid triticale*. *Triticale*. Springer International Publishing Switzerland.

Gowen, A.A., O'Donnell, C.P., Cullen, P.J., Downey, G. & Frias, J.M. (2007). Hyperspectral imaging - an emerging process analytical tool for food quality and safety control. *Trends in Food Science and Technology*, **18**, 590–598.

Greenaway, W.T. (1969). A Wheat Hardness Index. *Cereal Sci Today*, **14**, 4–7.

Greenwell, P. & Schofield, J. (1986). A starch granule protein associated with endosperm softness in wheat. *Cereal chemistry*, **63**, 379–380.

Greenwell, P. & Schofield, J.D. (1989). The chemical basis of grain hardness and softness. *H. Salovaara (Ed.)*

Wheat End-Use Properties, Proceedings ICC '89 Symposium (Lahti, Finland), 59–72.

Herschel, W. (1832). Investigation of the powers of the prismatic colours to heat and illuminate objects; with remarks, that prove the different refrangibility of radiant heat. To which is added, an inquiry into the method of viewing the sun advantageously, with telescopes of. *Abstracts of the Papers Printed in the Philosophical Transactions of the Royal Society of London*, **1**, 20–21.

Hruschka, W.R. & Norris, K.H. (1982). Least-Squares Curve Fitting of near Infrared Spectra Predicts Protein and Moisture Content of Ground Wheat. *Applied Spectroscopy*, **36**, 261–265.

Ibrahim, A., Varga, A.C., Jolánkai, M. & Safranyik, F. (2018). Applying infrared technique as a nondestructive method to assess wheat applying infrared technique as a nondestructive method to assess wheat grain hardness. **3**, 100-107.

Igne, B., Gibson, L.R., Rippke, G.R., Schwarte, A. & Hurburgh, C.R. (2007). Triticale moisture and protein content prediction by near-infrared spectroscopy (NIRS). *Cereal Chemistry*, **84**, 328–330.

J., B.R. & Dyck, P.L. (1975). Relation of several quality characteristics to hardness in two spring wheat crosses. *Canadian Journal of Plant Science*, **55**, 625–627.

Jimenez, R., Molina, L., Zarei, I., Lapis, J.R., Chavez, R., Cuevas, R.P.O. & Sreenivasulu, N. (2019). Method development of near-infrared spectroscopy approaches for nondestructive and rapid estimation of total protein in brown rice flour. In: *Methods in Molecular Biology*. Pp. 109–135.

Jolly, C.J., Rahman, S., Kortt, A.A. & Higgins, T.J.V. (1993). Characterisation of the wheat Mr 15000 "grain-softness protein" and analysis of the relationship between its accumulation in the whole seed and grain softness. *Theoretical and Applied Genetics*, **86**, 589–597.

Kays, S.E., Barton, F.E. & Windham, W.R. (2000). Predicting protein content by near infrared reflectance spectroscopy in diverse cereal food products. *Journal of Near Infrared Spectroscopy*, **8**, 35–43.

Khan, K. & Shewry, P.R. (2009). *Wheat: Chemistry and technology*. *A Companion to the Philosophy of Technology*. Fourth Edi. St. Paul, Minnesota: AACC International.

Khateeb, W. Al, Shalabi, A. Al, Schroeder, D. & Musallam, I. (2017). Phenotypic and molecular variation in drought tolerance of Jordanian durum wheat (Triticum durum Desf.) landraces. *Physiology and Molecular Biology of Plants*, **23**, 311–319.

Kirchler, C.G., Pezzei, C.K., Beć, K.B., Henn, R., Ishigaki, M., Ozaki, Y. & Huck, C.W. (2017). Critical evaluation of NIR and ATR-IR spectroscopic quantifications of rosmarinic acid in rosmarini folium supported by quantum chemical calculations. *Planta Medica*, **83**, 1076–1084.

Lachman, J., Martinek, P., Kotíková, Z., Orsák, M. & Šulc, M. (2017). Genetics and chemistry of pigments in wheat grain – A review. *Journal of Cereal Science*, **74**, 145–154.

Lai, F.S., Rousser, R., Brabec, D. & Pomeranz, Y. (1985). Determination of hardness in wheat mixtures .2. apparatus for automated measurement of hardness of single kernels. *Cereal Chemistry*, **62**, 178–184.

Li, C.Y., Li, W.H., Lee, B., Laroche, A., Cao, L.P. & Lu, Z.X. (2011). Morphological characterization of triticale starch granules during endosperm development and seed germination. *Canadian Journal of Plant Science*, **91**, 57–67.

Liu, Y., Pu, H. & Sun, D.W. (2017). Hyperspectral imaging technique for evaluating food quality and safety during various processes: A review of recent applications. *Trends in Food Science and Technology*, **69**, 25–35.

Maghirang, E.B. & Dowell, F.E. (2003). Hardness measurement of bulk wheat by single-kernel visible and near-infrared reflectance spectroscopy. *Cereal Chemistry*, **80**, 316–322.

Mahesh, S., Jayas, D.S., Paliwal, J. & White, N.D.G. (2014). Comparison of partial least squares regression (PLSR) and principal components regression (PCR) methods for protein and hardness predictions using the near-infrared (NIR) hyperspectral images of bulk samples of canadian wheat. *Food and Bioprocess Technology*, **8**, 31–40.

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: Non-destructive analysis of biological materials. *Chemical Society Reviews*, **43**, 8200–8214.

Manley, M., McGoverin, C.M., Snyders, F., Muller, N., Botes, W.C. & Fox, G.P. (2013). Prediction of triticale grain quality properties, based on both chemical and indirectly measured reference methods, using near-infrared spectroscopy. *Cereal Chemistry*, **90**, 540–545.

Manley, M., Toit, G. du & Geladi, P. (2011). Tracking diffusion of conditioning water in single wheat kernels of different hardnesses by near infrared hyperspectral imaging. *Analytica Chimica Acta*, **686**, 64–75.

Manley, M., Zyl, L. Van & Osborne, B.G. (2002). Using fourier transform near infrared spectroscopy in determining kernel hardness, protein and moisture content of whole wheat flour. *Journal of Near Infrared Spectroscopy*, **10**, 71–76.

Martin, C.R., Rousser, R. & Brabec, D.L. (1993). Martin_singlekernelsystem.pdf. *American Society of Agricultural Engineering*, **36**, 1399–1404.

Mattern, P.J. (1988). Wheat hardness: a microscopic classification of individual grains. *Cereal Chemistry*, **65**, 312–315.

Mattern, P.J., Morris, R., Schmidt, J.W. & Johnson, V.A. (1973). Location of genes for kernel properties in the wheat cultivar ‘Cheyenne’ using chromosome substitution lines. *Proceedings of the 4th International Wheat Genetics Symposium*, 703–707.

McClure, W.F. (2003). 204 Years of near infrared technology: 1800-2003. *Journal of Near Infrared Spectroscopy*, **11**, 487–518.

Mcgoverin, C.M., Snyders, F., Muller, N., Botes, W., Fox, G. & Manley, M. (2011). A review of triticale uses and the effect of growth environment on grain quality. *Journal of the Science of Food and Agriculture*, **91**, 1155–1165.

Mergoum, M., Pfeiffer, W. H., Pe~na, R. J., Ammar, K., Rajaram, S. (2004). Triticale crop improvement: the CIMMYT programme. *Triticale improvement and production*. FAO.

Merwe, J.D. van der & Cloete, P.C. (2018). Financial impact of wheat quality standards on South African wheat producers: A dynamic linear programming (DLP) approach. *Development Southern Africa*, **35**, 53–69.

Miller, B.S., Afework, S., Hughes, J.W. & Pomeranz, Y. (1981). Wheat hardness: time required to grind wheat with the brabender automatic. micro hardness tester. *Journal of Food Science*, **46**, 1863-1865

Mohammadi, R. (2016). Efficiency of yield-based drought tolerance indices to identify tolerant genotypes in durum wheat. *Euphytica*, **211**, 71–89.

Monneveux, P., Zaharieva, M. & Rekika, D. (2000). The utilisation of triticum and aegilops species for the improvement of durum wheat. *Durum wheat improvement in the Mediterranean region: New challenges*, **81**, 71–81.

Morales-Sillero, A., Fernández Pierna, J.A., Sinnaeve, G., Dardenne, P. & Baeten, V. (2018). Quantification of protein in wheat using near infrared hyperspectral imaging: Performance comparison with conventional near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, **26**, 186–195.

Morris, C.F., Greenblatt, G.A., Bettge, A.D. & Malkawi, H.I. (1994). Isolation and characterization of multiple forms of friabilin. *Journal of Cereal Science*, **20**, 167–174.

Muhamad, I.I. & Campbell, G.M. (2004). Effects of kernel hardness and moisture content on wheat breakage in the single kernel characterisation system. *Innovative Food Science and Emerging Technologies*, **5**, 119–125.

Müller, J. (2017). Dumas or Kjeldahl for reference analysis? *Analytics beyond measure*, 1–5.

Munir, M.T., Wilson, D.I., Yu, W. & Young, B.R. (2018). An evaluation of hyperspectral imaging for characterising milk powders. *Journal of Food Engineering*, **221**, 1–10.

Næs, T. & Martens, H. (1988). Principal component regression in NIR analysis: Viewpoints, background details and selection of components. *Journal of Chemometrics*, **2**, 155–167.

Nielsen, J.P., Pedersen, D.K. & Munck, L. (2003). Development of nondestructive screening methods for single kernel characterization of wheat. *Cereal Chemistry*, **80**, 274–280.

Norris, K.H. (1996). History of NIR. *Journal of Near Infrared Spectroscopy*, **4**, 31–37.

Oda, S. (1994). Two-dimensional electrophoretic analysis of friabilin. *Cereal Chemistry*, **71**, 394–395.

Orman, B.A. & Schumann, R.A. (1991). Comparison of near-infrared spectroscopy calibration methods for the prediction of protein, oil, and starch in maize grain. *Journal of Agricultural and Food Chemistry*, **39**, 883–886.

Orth, R.A. & Shellenberger, J.A. (1988). Origin, production, and utilization of wheat. *Association of Cereal Chemists, Inc …*, 14.

Osborne, B.G. & Anderssen, R.S. (2003). Single-kernel characterization principles and applications. *Cereal Chemistry*.

Osborne, B.G., Douglas, S. & Fearn, T. (1981). Assessment of wheat grain texture by near infrared reflectance measurements on bühler-milled flour. *Journal of the Science of Food and Agriculture*, **32**, 200-202.

Osborne, T.B. (2011). The proteins of the wheat kernel. *The proteins of the wheat kernel.*, **84**.

Pasikatan, M.C. & Dowell, F.E. (2004). High-speed nir segregation of high- and low-protein single wheat seeds. *Cereal Chemistry*.

Pomeranz, Y., Bolling, H. & Zwingelberg, H. (1984). Wheat hardness and baking properties of wheat flours. *Journal of Cereal Science*, **2**, 137–143.

Pomeranz, Y. & Williams, P.C. (1990). Wheat hardness. Its genetic, structural, and biochemical background, measurement, and significance. *Advances in Cereal Science and Technology*, **10**, 471–544.

Porep, J.U., Kammerer, D.R. & Carle, R. (2015). On-line application of near infrared (NIR) spectroscopy in food production. *Trends in Food Science and Technology*, **46**, 211–230.

Qu, J.H., Liu, D., Cheng, J.H., Sun, D.W., Ma, J., Pu, H. & Zeng, X.A. (2015). Applications of Near-infrared Spectroscopy in Food Safety Evaluation and Control: A Review of Recent Research Advances. *Critical Reviews in Food Science and Nutrition*, **55**, 1939–1954.

Quayson, E.T., Atwell, W., Morris, C.F. & Marti, A. (2016). Empirical rheology and pasting properties of soft-textured durum wheat (Triticum turgidum ssp. durum) and hard-textured common wheat (T. aestivum). *Journal of Cereal Science*, **69**, 252–258.

Ravikanth, L., Singh, C.B., Jayas, D.S. & White, N.D.G. (2016). Performance evaluation of a model for the classification of contaminants from wheat using near-infrared hyperspectral imaging. *Biosystems Engineering*, **147**, 248–258.

Rinnan, Å. (2014). Pre-processing in vibrational spectroscopy-when, why and how. *Analytical Methods*. **6**, 7124-7129

Rinnan, Å., Berg, F. van den & Engelsen, S.B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends in Analytical Chemistry*, **28**, 1201–1222.

S. R. Delwiche. (1993). Measurement of Single-kernel Wheat Hardness Using Near-infrared Transmittance. *Transactions of the ASAE*, **36**, 1431–1437.

Saleh, M.I., Meullenet, J.F. & Siebenmorgen, T.J. (2008). Development and validation of prediction models for rice surface lipid content and color parameters using near-infrared spectroscopy: A basis for predicting rice degree of milling. *Cereal Chemistry*, **85**, 787–791.

Salmanowicz, B.P., Adamski, T., Surma, M., Kaczmarek, Z., Karolina, K., Kuczyńska, A., Banaszak, Z., Ługowska, B., Majcher, M. & Obuchowski, W. (2012). The relationship between grain hardness, dough mixing parameters and bread-making quality in winter wheat. *International Journal of Molecular Sciences*, **13**, 4186–4201.

Sampson, D.R., Flynn, D.W. & Jui, P. (1983). Genetic studies on kernel hardness in wheat using grinding time and near infrared reflectance spectroscopy. *Canadian Journal of Plant Science*, **63**, 825-832.

Sendin, K., Williams, P.J. & Manley, M. (2018). Near infrared hyperspectral imaging in quality and safety evaluation of cereals. *Critical Reviews in Food Science and Nutrition*, **58**, 575–590.

Shewry, P.R. (2009). Wheat. *Journal of Experimental Botany*, **60**, 1537–1553.

Shewry, P.R. (2018). Do ancient types of wheat have health benefits compared with modern bread wheat? *Journal of Cereal Science*, **79**, 469–476.

Shewry, P.R. & Hey, S. (2015). Do “ancient” wheat species differ from modern bread wheat in their contents of bioactive components? *Journal of Cereal Science*, **65**, 236–243.

Stenvert, N.L. (1974). Grinding resistance, a simple measure of wheat hardness. *Flour Anim Feed Milling*, **12**, 24–26.

Stenvert, N.L. & Kingswood, K. (1977). The influence of the physical structure of the protein matrix on wheat hardness. *Journal of the Science of Food and Agriculture*, **28**, 11–19.

Symes, K.J. (1965). The inheritance of grain hardness in wheat as measured by the particle size index. *Australian Journal of Agricultural Research*, **16**, 113–123.

Tanno, K.I. & Willcox, G. (2006). How fast was wild wheat domesticated? *Science*, **311**, 1886.

Turnbull, K.M., Marion, D., Gaborit, T., Appels, R. & Rahman, S. (2003). Early expression of grain hardness in the developing wheat endosperm. *Planta*, **216**, 699–706.

Varughese, G., Pfeiffer, W.H. & Peña, R.J. (1996). Triticale: A successful alternative crop (Part 1). *Cereal Foods World*, **41**, 474–482.

Vines, L.L., Kays, S.E. & Koehler, P.E. (2005). Near-infrared reflectance model for the rapid prediction of total fat in cereal foods. *Journal of Agricultural and Food Chemistry*, **53**, 1550–1555.

Wang, H.L., Wan, X.Y., Bi, J.C., Wang, J.K., Jiang, L., Chen, L.M., Zhai, H.Q. & Wan, J.M. (2006). Quantitative analysis of fat content in rice by near-infrared spectroscopy technique. *Cereal Chemistry*, **83**, 402–406.

Wang, L., Sun, D.W., Pu, H. & Cheng, J.H. (2017). Quality analysis, classification, and authentication of liquid foods by near-infrared spectroscopy: A review of recent research developments. *Critical Reviews in Food Science and Nutrition*, **57**, 1524–1538.

Weegels, P.L., Hamer, R.J. & Schofield, J.D. (1996). Functional properties of wheat glutenin. *Journal of Cereal Science*, **23**, 1-17.

Wesley, I.J., Ruggiero, K., Osborne, B.G. & Anderssen, R.S. (2005). The challenge of single estimates in near infrared calibration and prediction: The measurement of durum vitreousness using receival instruments. *Journal of Near Infrared Spectroscopy*, **13**, 333–338.

Williams, P. (1991). Prediction of wheat kernel texture in whole grains by near infrared trasmittance. *Cereal Chemistry*, **68**, 112–114.

Williams, P., Antoniszyn, J. & Manley, M. (2019). *Near-infrared Technology: Getting the best out of light*. *Near-infrared Technology: Getting the best out of light*. AFRICAN SUN MeDIA.

Williams, P.C. & Sobering, D.C. (1993). Comparison of Commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. *Journal of Near Infrared Spectroscopy*, **1**, 25–32.

Williams, P.C., Stevenson, S.G., Starkey, P.M. & Hawtin, G.C. (1978). The application of near infrared reflectance spectroscopy to protein-testing in pulse breeding programmes. *Journal of the Science of Food and Agriculture*, **29**, 285–292.

Windham, W.R., Kays, S.E. & Barton, F.E. (1997). Effect of cereal product residual moisture content on total dietary fiber determined by near-infrared reflectance spectroscopy. *Journal of Agricultural and Food Chemistry*, **45**, 140–144.

Wold, S., Sjöström, M. & Eriksson, L. (2001). *PLS-regression: A basic tool of chemometrics*. *Chemometrics and Intelligent Laboratory Systems*, **58**, 109-130.

Wrigley, C., Corke, H., Seetharaman, K. & Faubion, J. (2016). *Encyclopedia of food grains*. *encyclopedia of food grains*. 2nd edn. Amsterdam: Elsevier.

Wu, D. & Sun, D.W. (2013). Advanced applications of hyperspectral imaging technology for food quality and safety analysis and assessment: A review - Part I: Fundamentals. *Innovative Food Science and Emerging Technologies*, **19**, 1–14.

Zhu, F. (2018). Triticale: Nutritional composition and food uses. *Food Chemistry*, **241**, 468–479.

Chapter 3: Materials and methods

3.1 Wheat and triticale samples

Sound, whole grain wheat and triticale samples were obtained from the Stellenbosch University Plant Breeding Laboratory (SU-PBL; Stellenbosch University, Stellenbosch, South Africa). The samples comprised 20 winter wheat and 20 triticale breeding lines from the 2018 harvest year. The wheat was planted with three replicates across three growing regions in the Western Cape of South Africa, i.e. Tygerhoek, Napier and Vredenburg. In total, 180 wheat and 177 triticale samples were obtained for bulk sample model development. For the single kernel model development, 39 kernels were randomly selected from each of these bulk samples resulting in 7020 wheat and 6903 triticale single kernels.

3.2 Protein and moisture content and kernel hardness determination

The bulk samples (5 g) were milled using a Retsch centrifugal hammer mill (Retsch GmbH, Haan, Germany) fitted with a 0.5 mm sieve. Moisture content was determined in duplicate using a TGM800 automated thermogravimetric moisture determinator (LECO Corporation, St. Joseph, Michigan, USA) in accordance with AACC International Approved method 44-15.02 (AACC Approved Methods of Analysis, 1999a). Total protein content was determined in duplicate by the Dumas combustion method according to AACC International Approved Method 46-30.01 (AACC Approved Methods of Analysis, 1999b) with a Gerhardt Dumatherm DT N40+ 14-0000 (Gerhardt Analytical Systems, Königswinter, Germany). Sample kernel hardness was determined using a Perten SKCS 4100 Single Kernel Characterisation System (SKCS) according to AACC International Approved Method 55-31.01 (AACC Approved Methods of Analysis, 1999c)

3.3 Near-infrared hyperspectral image system setup and image acquisition

NIR reflectance images of the wheat kernels were obtained using a HySpex SWIR-384 (HySpex, Skedsmokorset, Norway) camera (Fig. 3.1). The spectral range for the camera was 930-2500 nm with 384 spatial pixels and 288 spectral channels with a spectral interval of 6 nm and spatial resolution of 53 µm. The optical sensor was an HgCdTe detector with built-in cooling to 150 Kelvin and a maximum frame rate of 400 frames per second (fps). Images were acquired using an 84 mm focal length lens at a working distance of 0.3 m and a field-of-view of 20 mm resulting in a pixel size of 52.9 µm. The light source consisted of two halogen direct current (DC) linear lamps with a wavelength range of 400-2500 nm and power consumption of 150 W each – mounted 20 cm above the translation stage and angled at 54 degrees. The imaging setup was equipped with a translation stage and constant feed rate was set at 50 mm/s. Grey and internal black reference standards were taken every 30 min and after every translation movement, respectively. The grey reference standard (Zenith Polymer® Reflectance Standards) was a 50% diffuse reflectance polytetrafluoroethylene (PTFE) standard, with constant reflection over the 250-2450 nm wavelength range. The sensor integration time was set at 2900 µs.

Wheat samples were placed on a specially designed nylon tray (Fig. 3.2), designed using AutoCAD Mechanical, 2018 (Autodesk®, Mill Valley, California, USA) and made black. The tray was designed and produced with single kernel size cut outs to enable imaging of 39 kernels (3 parallel rows of 13 kernels each) of each of seven samples (n=273) simultaneously. This allowed for assigning coordinates to the single kernels and ensured neat uncluttered images.

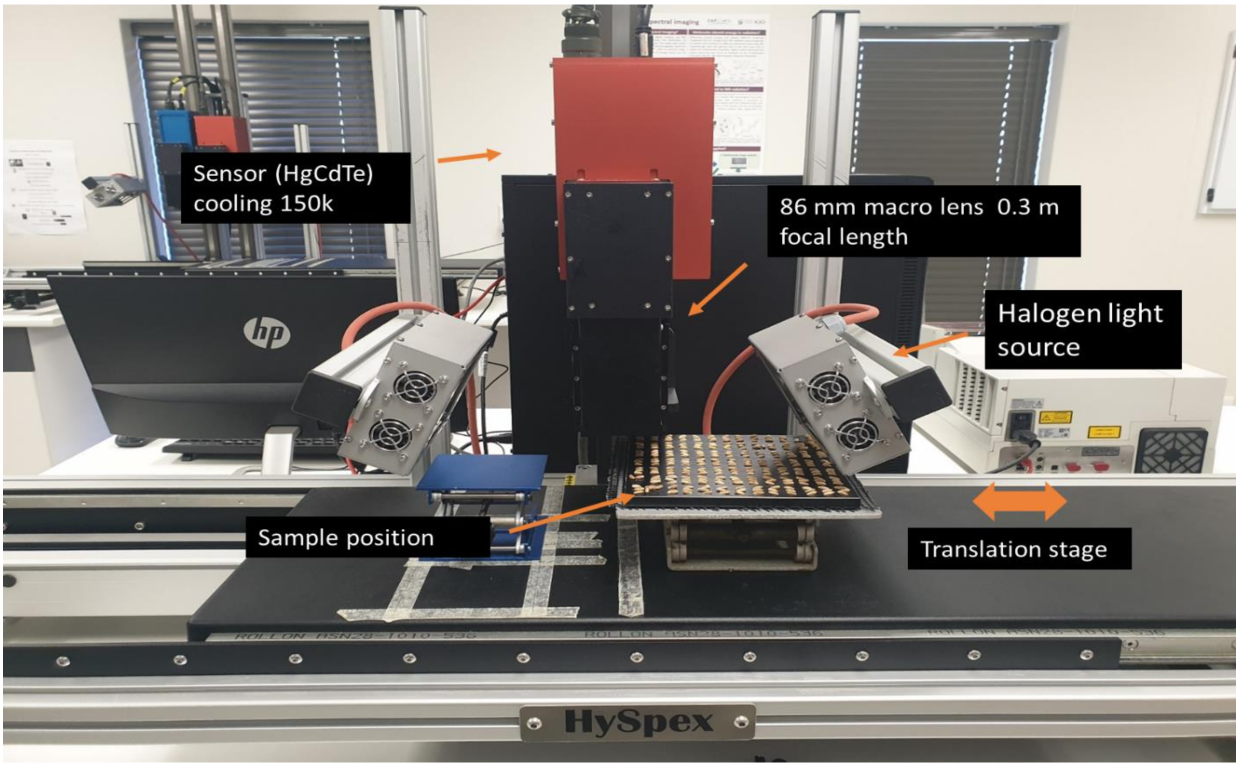

Figure 3.1 Near-infrared hyperspectral imaging setup with the HySpex SWIR 384 camera equipped with a HgCdTe sensor, two halogen light sources and a translation stage.

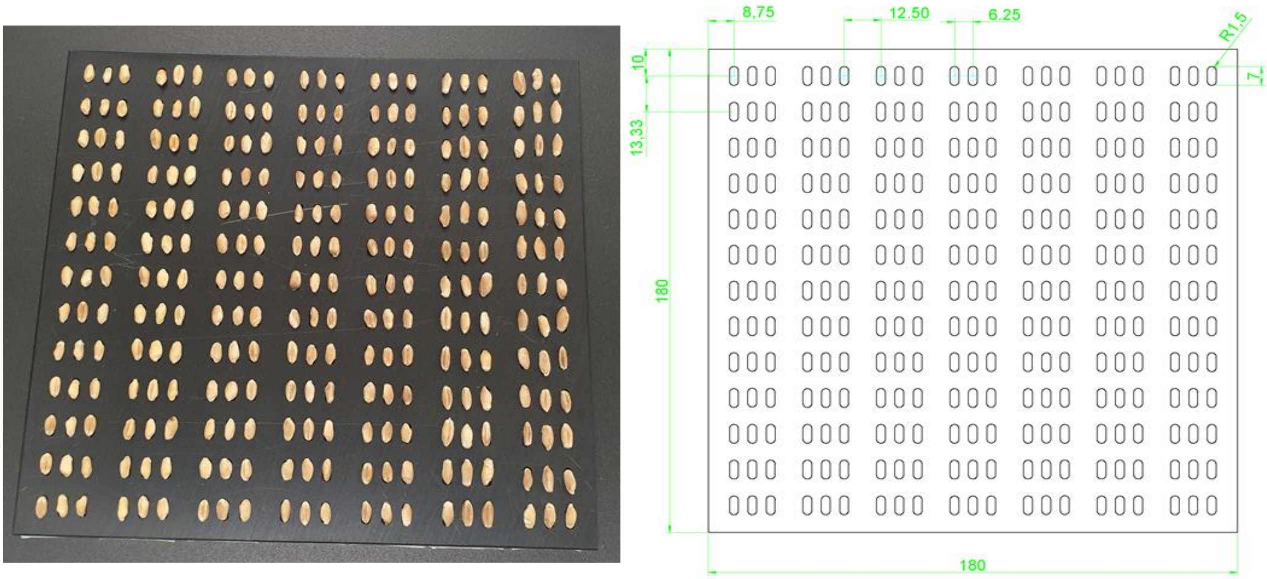

Figure 3.2 The black nylon tray used for sample presentation for spectral imaging allowing seven sets of 39 single kernels per sample to be imaged simultaneously.

3.4 Hyperspectral image analysis

After image acquisition, the spectral images were converted from reflectance to pseudo absorbance (Sendin *et al*., 2018) using the Evince v.2.7.0 (Prediktera, Umeå, Sweden) spectral image analysis software. Principal component analysis (PCA) was applied to the images and three principal components (PCs) were calculated. The background, dead pixels, shading and outlier pixels were removed from the data set using PC scores images and scores plots interactively. Objects-of-interest (single kernels) were identified and the average spectrum for each kernel obtained. A single spectrum for each bulk sample comprising 39 single kernels was also determined.

3.5 Partial least squares regression

The bulk sample and single kernel spectra for both wheat and triticale were further analysed using Matlab R2018b (MathWorks, Natick, Massachusetts, USA) and PLS-Toolbox (Eigenvector Research Inc, Manson, WA, USA). The spectra were truncated to 1100-2096 nm in order to reduce noise in the extremes of the spectra. PCA was performed on mean-centred average spectra of the bulk wheat, triticale and combined data sets, in order to detect outliers and groupings. Training and validation sets were selected using the DUPLEX algorithm at a 30% threshold (Snee, 1977). The number of samples included in the training and validation sets for the bulk and single kernel data sets is shown in Table 3.1.

Table 3.1 Number of samples in training and validation sets for bulk and single kernel wheat, triticale and combined data sets selected using the DUPLEX algorithm

	Bulk data sets			**Single kernel data sets**		
	Wheat	**Triticale**	**Combined**	**Wheat**	**Triticale**	**Combined**
Training set (70%)	126	124	250	4914	4833	9747
Validation set (30%)	54	53	107	2106	2070	4176

Pre-processing techniques evaluated included standard normal variate (SNV), detrend (DT) (Barnes *et al.*, 1989), mean centring (MC), orthogonal signal correction (OSC) (Sjöblom *et al.*, 1998), Savitzky-Golay second derivative (3rd order polynomial, 15 points) and first derivative (2nd order polynomial, 15 points) (Savitzky and Golay, 1964) and also generalised least squares (GLS) (Buse, 1973).

Calibration models were developed using the partial least squares (PLS) regression. Two PLS algorithms were evaluated, i.e. Straightforward Implementation of a statistically inspired Modification of the PLS method (SIMPLS) (de Jong, 1993) and robust-PLS (RSIMPLS) (Hubert and Branden, 2003). Single kernel outliers were removed using the robust-PLS algorithm and by evaluation of their predicted vs. measured Y residuals to manually remove outliers. Cross-validation was performed on the training set, to determine the optimum number of latent variables (LV), using venetian blinds with 14 splits and 5 samples per split for the bulk data set models and with 20 splits and 5 samples per split for the single kernel data set models. Calibration and prediction accuracies were evaluated by means of root mean square error of calibration (RMSEC), -cross-validation (RMSECV) and -prediction (RMSEP). Also, the coefficient of determination (R^2) for calibration (R^2_{cal}), cross-validation (R^2_{CV}) and prediction (R^2_{pred}) was taken into account.

An independent test set was obtained, and the models acquired by the SIMPLS and RSIMPLS methods were tested for protein content prediction accuracy. Wheat (76) and triticale (74) single kernels were selected at random from the sample set and the kernels were imaged and processed in the same manner as for the calibration set. The kernels were than individually analysed for protein content by the Dumas combustion method in order to obtain the reference. Lastly the spectral and reference data were introduced to the models as a test set in order to truly test the models performance. The experimental layout is summarised in Figure. 3.3 by means of a flow diagram.

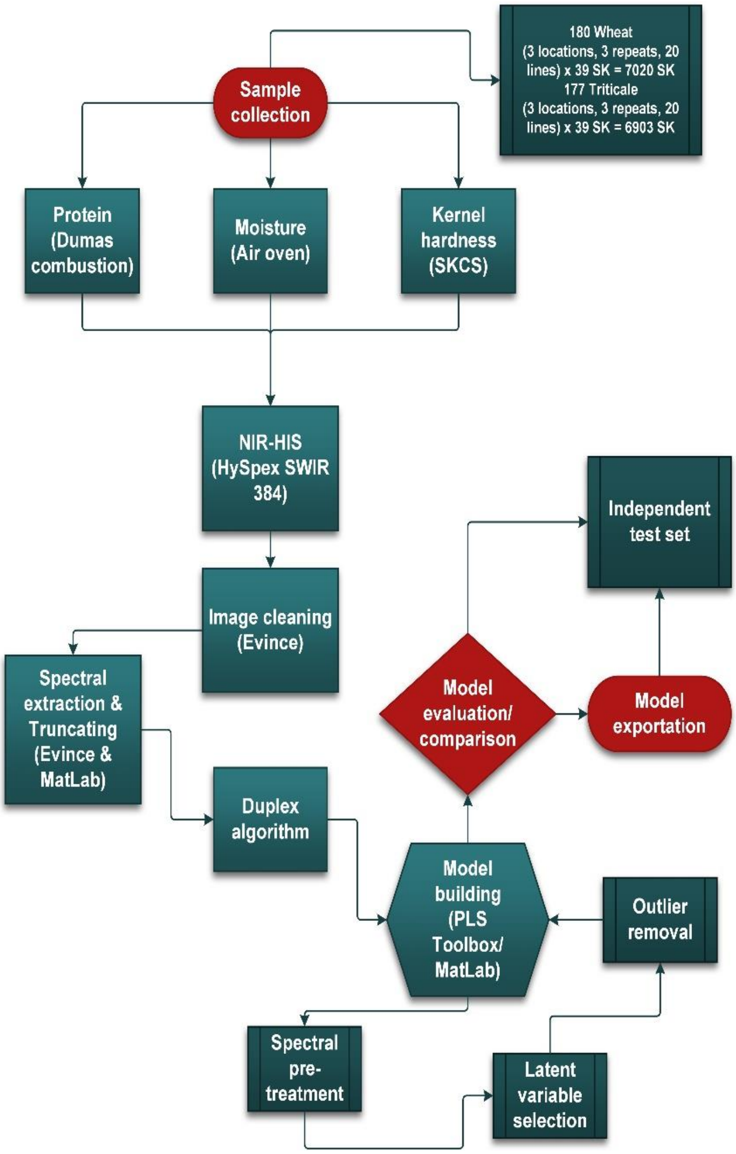
180 Wheat
(3 locations, 3 repeats, 20 lines) x 39 SK = 7020 SK
177 Triticale
(3 locations, 3 repeats, 20 lines) x 39 SK = 6903 SK
Sample collection
Protein (Dumas combustion)
Moisture (Air oven)
Kernel hardness (SKCS)
NIR-HIS (HySpex SWIR 384)
Independent test set
Image cleaning (Evince)
Spectral extraction & Truncating (Evince & MatLab)
Model evaluation/ comparison
Model exportation
Duplex algorithm
Model building (PLS Toolbox/ MatLab)
Outlier removal
Spectral pre-treatment
Latent variable selection

Figure 3.3 A flow diagram summarising the methodology used to build and evaluate models for wheat and triticale protein and moisture content and also kernel hardness prediction.

3.6 References

AACC Approved Methods of Analysis. (1999a). Method 44-15.02, Moisture-air-oven methods. 11th edn. St. Paul, MN, U.S.A.: Cereals & Grains Association.

AACC Approved Methods of Analysis. (1999b). Method 46-30.01, Crude protein-combustion method. 11th edn. St. Paul, MN, U.S.A.: Cereals & Grains Association.

AACC Approved Methods of Analysis. (1999c). Method 55-31.01. Single kernel characterization system (**skcs**) for wheat kernel texture. 11th edn. St. Paul, MN, U.S.A.: Cereals & Grains Association.

Barnes, R.J., Dhanoa, M.S. & Lister, S.J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, **43**, 772–777.

Buse, A. (1973). Goodness of fit in generalized least squares estimation. *The American Statistician*, **27**, 106-108.

Hubert, M. & Branden, K. Vanden. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*, **17**, 537–549.

Jong, S. de. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, **18**, 251–263.

Savitzky, A. & Golay, M.J.E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, **36**, 1627–1639.

Sjöblom, J., Svensson, O., Josefson, M., Kullberg, H. & Wold, S. (1998). An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. In: *Chemometrics and Intelligent Laboratory Systems*. Pp. 229–244. Elsevier.

Snee, R.D. (1977). Validation of regression models: methods and examples. *Technometrics*, **19**, 415–428.

Chapter 4: Results and discussion

4.1 Exploratory analysis

Figure 4.1 shows the average spectra of the bulk wheat, triticale and combined data sets as well as the PCA plots of the mean-centred data. A separation between the three localities, Vredenburg, Napier and Tygerhoek was observed in the direction of PC1 from left to right for both wheat (Fig. 4.1a) and triticale (Fig. 4.1b). The PCA plot of the combined data set showed no clustering, indicating that it was suitable to combine the spectra of these two grains to develop a single calibration model (Fig. 4.1c). After pre-processing with SNV, DT and 2nd derivative, some spectra showed a distinct spectral protuberance at *ca.* 1550-1650 nm. These spectra were identified in a PC1 vs. PC2 plot as those of wheat samples from a single experimental plot from Tygerhoek (Fig. 4.2). The spectral differences were emphasised by the pre-treatment techniques. SNV corrects for standard offset in absorbance, DT corrects for baseline shift across the variables and 2nd derivative emphasises the spectral differences. As neither moisture nor protein absorb at 1550–1650 nm (Williams *et al.*, 2019), these spectra were not removed from the data set. The protuberance could have been due to light scattering related to these samples and not effectively corrected for by the pre-treatment techniques.

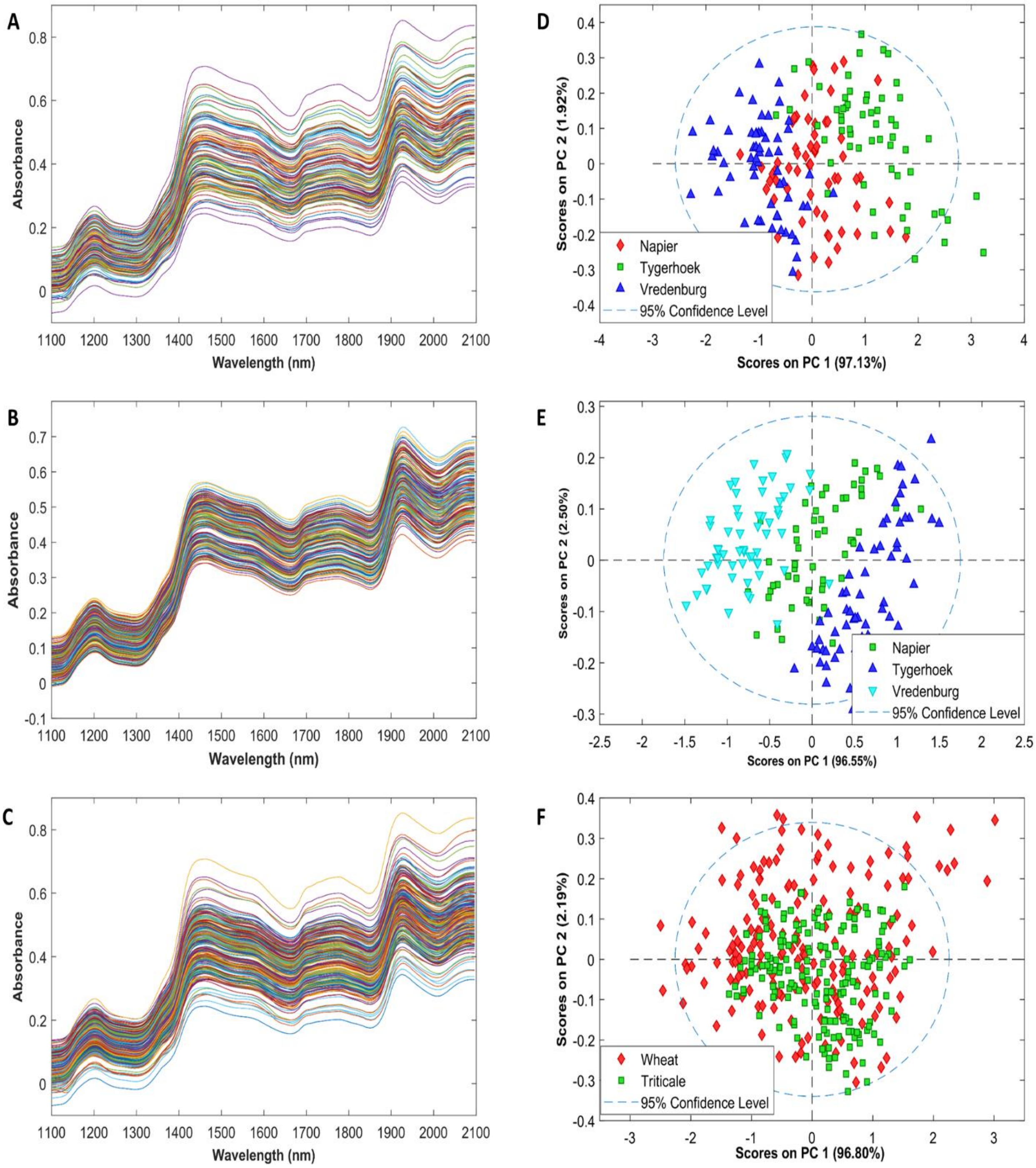

Figure 4.1 Whole kernel, raw average spectra of the (a) wheat, (b) triticale and (c) wheat and triticale combined data sets. PCA plots of PC1 vs. PC2 for the corresponding mean-centred spectral data with score values coloured based on location (Napier, Tygerhoek and Vredenburg) for (d) wheat, (e) triticale and on the (f) type of grain (wheat and triticale) in the combined data set.

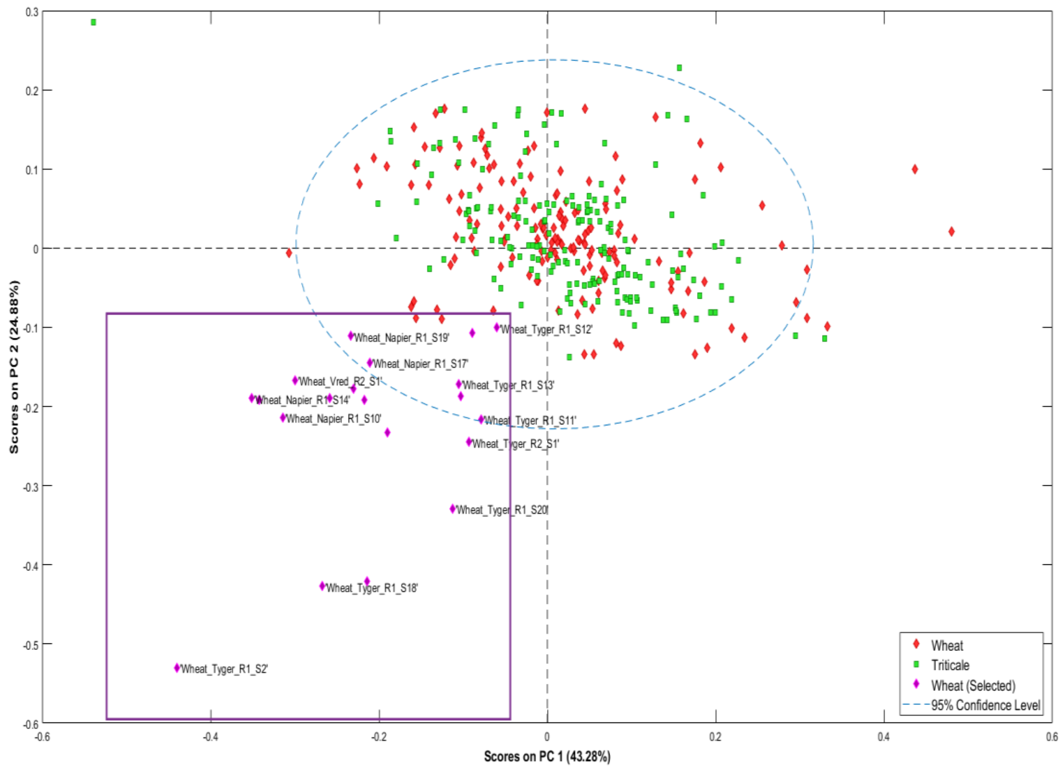

Figure 1.2. PCA plot of PC1 vs. PC2 of spectra from the combined data set pre-treated with SNV, DT and 2nd derivative, showing distinct clustering of wheat samples from an experimental plot mainly from Tygerhoek.

Descriptive statistics for protein and moisture content of the wheat, triticale and combined data sets are shown in Table 4.1 and the distributions are shown in histograms (Fig. 4.3). The histograms displayed Gaussian distribution for all data sets. This illustrated over representation of reference data around the 50% confidence interval whereas an equal distribution of the data across the entire range would be ideal (Williams *et al*., 2019). The training and validation sets were selected using the DUPLEX algorithm. The data range for the validation set falls within that of the training set.

Table 4.1 Descriptive statistics for protein and moisture content (%) of wheat and triticale samples

	Protein content (%)			**Moisture content (%)**		
	Wheat	**Triticale**	**Combined**	**Wheat**	**Triticale**	**Combined**
		Training set			**Training set**	
Mean	12.02	11.00	11.54	11.92	12.57	12.29
SD	1.05	1.21	1.25	1.03	1.10	1.06
Min	9.57	7.41	7.41	9.89	10.40	9.94
Max	14.66	14.66	14.66	13.40	14.40	14.40
Median	12.10	11.01	11.60	12.15	12.90	12.30

		Validation set			Validation set	
Mean	11.76	11.12	11.40	12.12	12.58	12.25
SD	1.02	1.21	1.12	0.97	1.04	1.14
Min	9.25	8.21	9.25	10.10	10.60	9.89
Max	13.33	13.91	13.91	13.20	13.80	14.00
Median	11.94	11.19	11.32	12.25	13.02	12.56

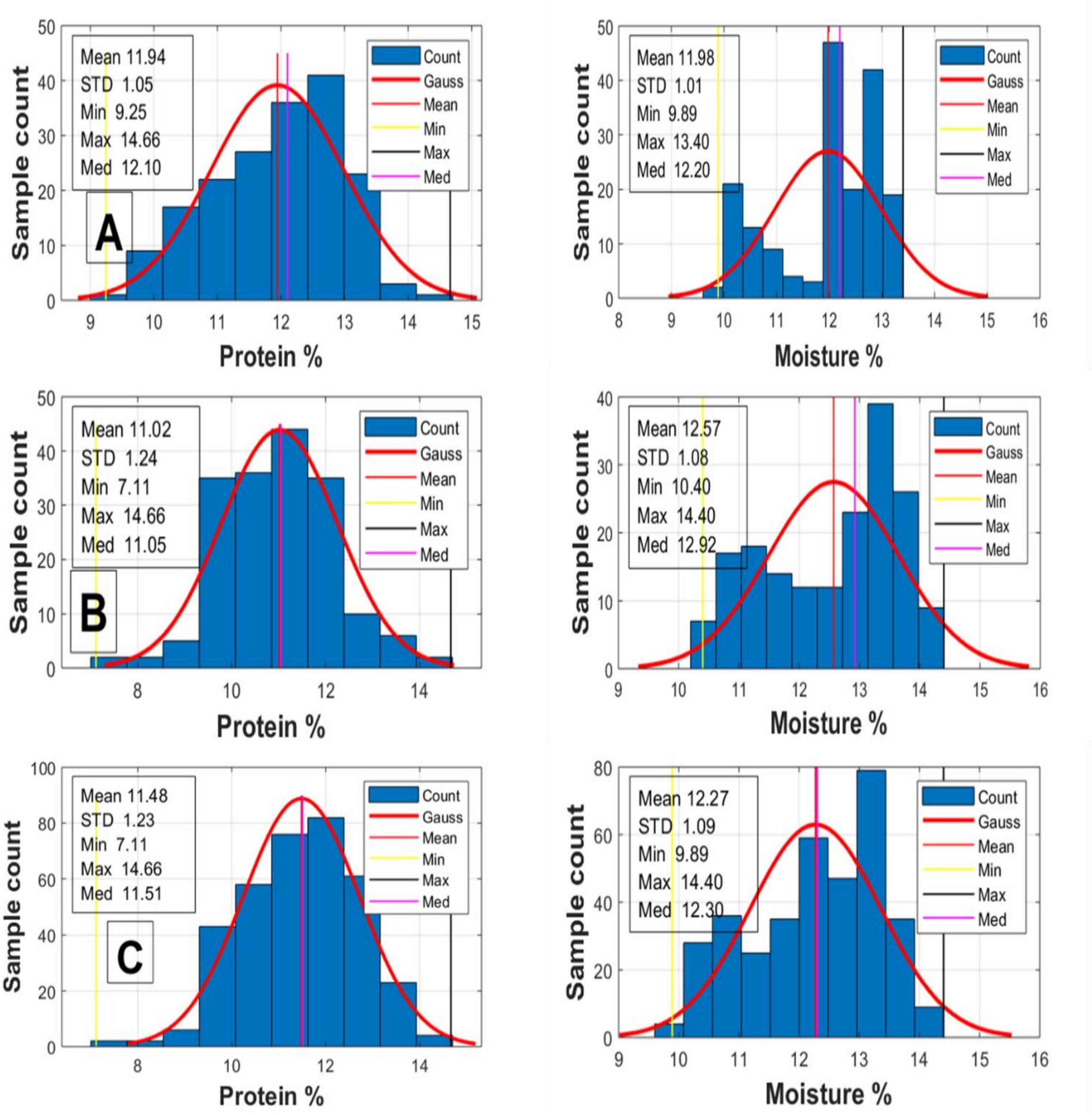

Figure 4.3 Histograms illustrating distribution of reference data, i.e. protein content for (a) wheat, (b) triticale and (c) wheat and triticale combined and moisture content for (c) wheat, (d) triticale and (e) wheat and triticale combined.

4.2 Bulk wheat and triticale PLS regression models

4.2.1 Protein content

The best wheat protein content PLS regression model was obtained with 2nd derivative pre-treated spectra (Table 4.2) with an RMSEP of 0.37% and R^2_P of 0.87. For triticale, the best model was obtained with the same pre-treatment resulting in an RMSEP of 0.53% and R^2_P 0.81. The best model for the combined data set was obtained with a combination of SNV, DT and 2nd derivative (RMSEP of 0.41% and R^2_P of 0.88). The results are comparable to that of Manley *et al.*, (2002), where a RMSEP of 1.16% with a R^2 of 0.81 was obtained.

Figure 4.4a shows the average pre-processed spectra of the combined wheat and triticale data set. The variable important in projection (VIP) scores plot for the combined data set protein prediction model (Fig. 4.4b) shows the important variables attributed to protein at 1430 (N-H 1st overtone) and 2000 (N-H combination nm). This is confirmed in the latent variable plot for latent variable (LV) 1, LV2 and LV11 (Fig. 4.4c).

Figure 4.5 shows the RMSEC, RMSECV and RMSEP values for increasing number of latent variables (LV's) for wheat and triticale combined data set. Overfitting of the model is apparent after 11 LV's, as the difference between RMSEC and RMSECV increases. If a model has been overfitted the model would not add to prediction accuracy, but would rather be detrimental to model performance. Figure 4.6 shows the protein content predicted vs. measured plot for the combined data set. Calibration samples at the higher and lower protein content values with large residuals could be considered as outliers, however no samples were removed due to the relatively small sample set.

Table 4.2 Calibration and validation statistics for protein content PLS regression models for bulk wheat, triticale and combined data sets using different pre-processing methods. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set										
#	**Preprocessing**	**XT**	**XV**	**LV**	**RMSEC**	**R^2_C**	**RMSECV**	**R^2_{CV}**	**RMSEP**	**R^2_P**
1	SNV, 2nd der (order: 3, window: 15 pt)	126	54	10	0.33	0.90	0.43	0.83	0.39	0.86
2	SNV	126	54	11	0.40	0.86	0.56	0.73	0.45	0.81
3	SNV, DT	126	54	15	0.30	0.92	0.40	0.86	0.41	0.84
4	Mean-centred, SNV, DT	126	54	10	0.61	0.66	0.72	0.54	0.68	0.57
5	OSC	126	54	16	0.30	0.92	0.42	0.84	0.40	0.84
6	**2nd der (order: 3, window: 15 pt)**	**126**	**54**	**10**	**0.34**	**0.90**	**0.43**	**0.83**	**0.37**	**0.87**
7	None	126	54	16	0.32	0.91	0.45	0.82	0.40	0.84
CV	venetian blinds w/ 14 splits and 5 samples per split									
Triticale data set										
#	**Preprocessing**	**XT**	**XV**	**LV**	**RMSEC**	**R^2_C**	**RMSECV**	**R^2_{CV}**	**RMSEP**	**R^2_P**
1	SNV, DT, 2nd der (order: 3, window: 15 pt)	123	54	10	0.55	0.79	0.67	0.70	0.54	0.80
2	SNV	123	54	13	0.55	0.79	0.69	0.68	0.58	0.78
3	SNV, DT	123	54	10	0.57	0.78	0.66	0.71	0.59	0.76
4	Mean-centred, SNV, DT	123	54	8	0.80	0.60	0.88	0.51	0.77	0.61
5	OSC	123	54	11	0.65	0.73	0.79	0.61	0.61	0.77
6	**2nd der (order: 3, window: 15 pt)**	**123**	**54**	**11**	**0.54**	**0.80**	**0.67**	**0.69**	**0.53**	**0.81**
7	None	123	54	13	0.55	0.79	0.69	0.68	0.58	0.78
CV	venetian blinds w/ 14 splits and 5 samples per split									
Combined data set										
#	**Preprocessing**	**XT**	**XV**	**LV**	**RMSEC**	**R^2_C**	**RMSECV**	**R^2_{CV}**	**RMSEP**	**R^2_P**
1	**SNV, DT, 2nd der (order: 3, window: 15 pt)**	**249**	**106**	**11**	**0.53**	**0.82**	**0.60**	**0.77**	**0.41**	**0.88**
2	SNV	249	106	15	0.51	0.84	0.60	0.77	0.47	0.83
3	SNV, DT	249	106	12	0.53	0.82	0.67	0.72	0.48	0.82
4	Mean Center, SNV, DT	249	106	10	0.73	0.65	0.82	0.56	0.71	0.61
5	OSC	249	106	12	0.53	0.82	0.62	0.75	0.50	0.81
6	2nd der (order: 3, window: 15 pt,)	249	106	11	0.49	0.84	0.58	0.78	0.44	0.85
7	None	249	106	13	0.57	0.79	0.64	0.74	0.51	0.81
CV	venetian blinds with 14 splits and 5 samples per split									

*Standard normal variate (SNV), Detrend (DT), orthogonal signal correction (OSC)

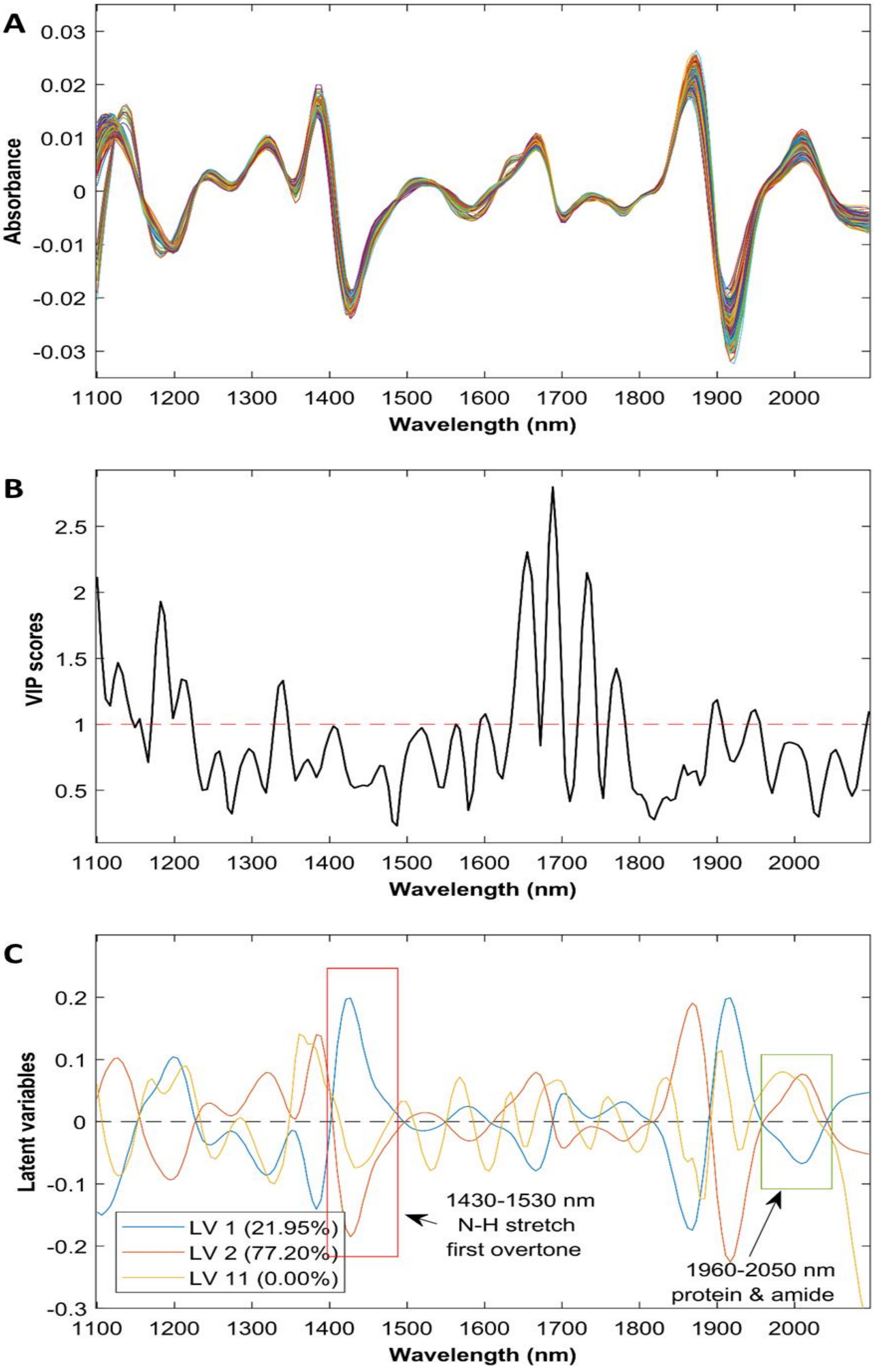

Figure 4.4 (a) Average pre-processed spectra of the combined data set for protein content prediction model 1 (Table 4.2) , (b) variable importance in projection and (c) latent variables for LV 1, LV 2 and LV 11.

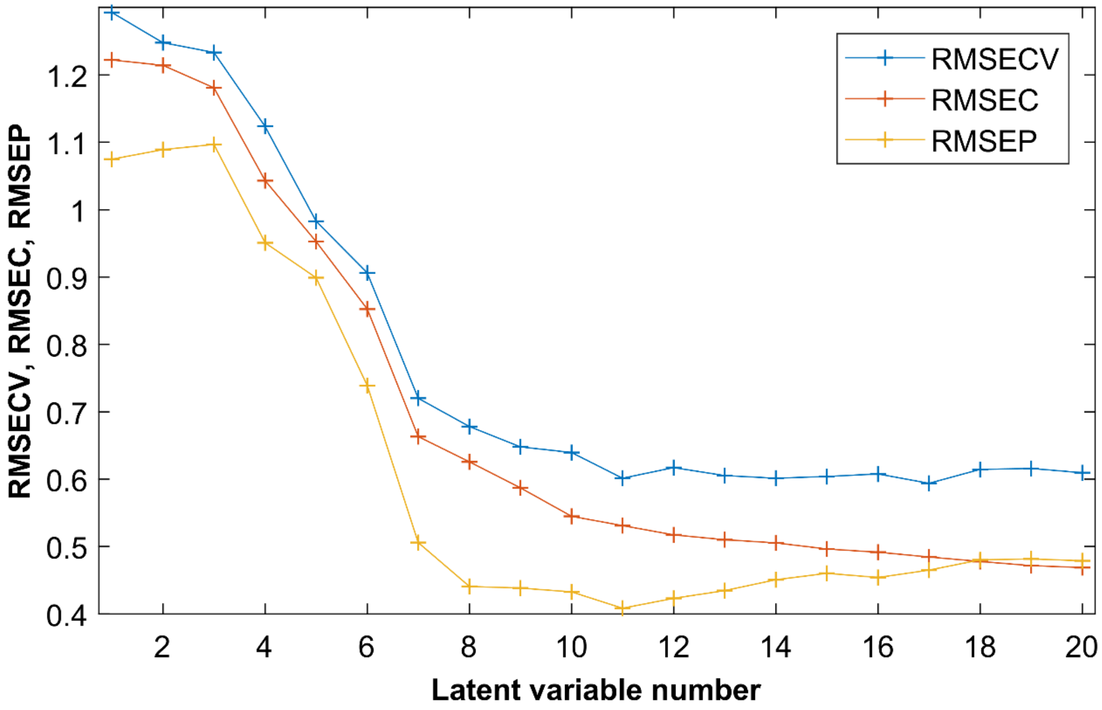

Figure 4.5 Latent variables vs. standard error of cross-validation, -calibration and -prediction for the combined data set protein prediction model 1 (Table 4.2.).

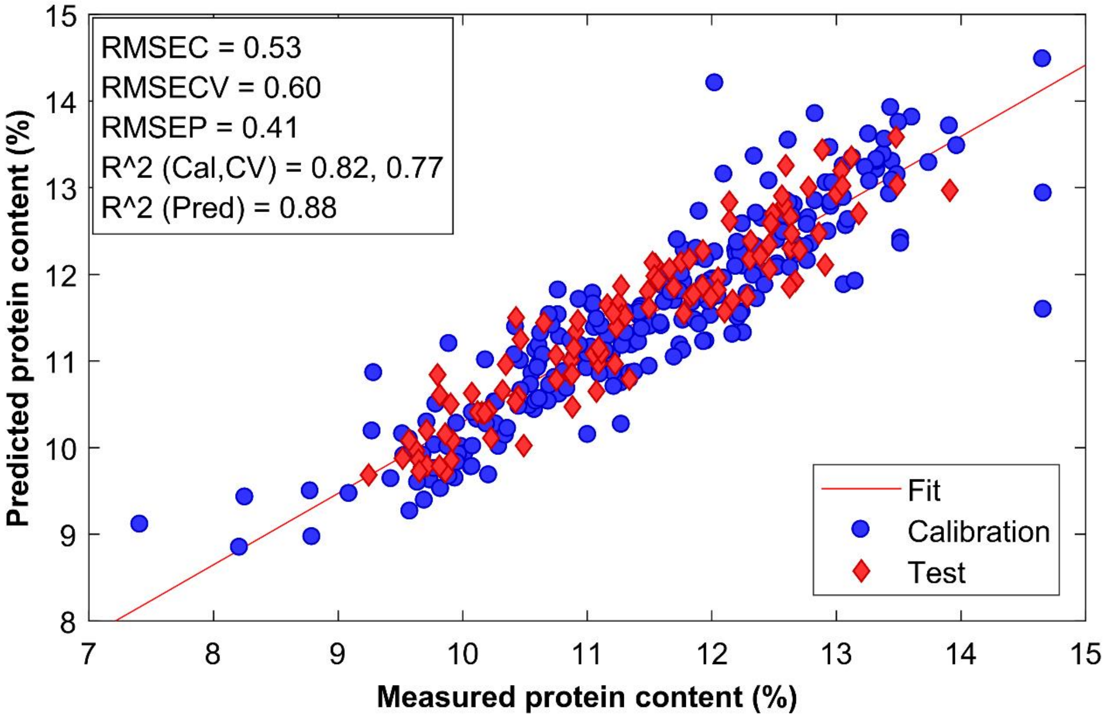

Figure 4.6 Measured vs. predicted protein content for PLS regression model 1 with SNV, DT and 2nd derivative pre-treatment for the combined wheat and triticale data set using 11 LV's and 249 samples in the training set and 106 in the validation set.

4.2.2 Moisture content

The best wheat moisture content PLS regression model (Table 4.3) was obtained with no pre-treatment with an RMSEP of 0.49% and R^2_P of 0.75. For triticale the spectra was pre-treated with SNV and 1st derivative (Table 4.3) resulting in an RMSEP of 0.36% and R^2_P of 0.88. The best model for the combined data set was obtained with a combination of SNV, DT and 2nd derivative (RMSEP of 0.49%; R^2_P of 0.82). This is highly comparable to previous studies done by Williams *et al.*, (1985); Manley *et al.*, (2002); Dowell *et al.*, (2006).

Figure 4.7a shows the average pre-processed spectra of the combined wheat and triticale data set pre-treated with SNV, DT and 2nd derivative used for the moisture content prediction. The VIP scores plot (Fig. 4.7b) shows that the variables of importance for moisture content prediction (1410-1450 and 1940 nm) contribute to moisture content prediction. This is confirmed in the latent variable plot for LV1, LV2 and LV12 (Fig. 4.7c).

Figure 4.8 shows the RMSEC, RMSECV and RMSEP values as a function of LV's for the combined data set. Model overfitting is apparent after 12 LV's, as the difference between RMSEC and RMSECV increases. Figure 4.9 shows the moisture content predicted vs. measured plot for the combined data set.

Table 3.3 Calibration and validation statistics for predicted moisture PLS regression models for bulk wheat, triticale and combined data sets using different pre-processing methods. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set

#	Preprocessing	XT	XV	LV	RMSEC	R2$_C$	RMSECV	R2$_{CV}$	RMSE$_P$	R2$_P$
1	SNV, DT	126	54	8	0.51	0.74	0.55	0.7	0.51	0.72
2	1st der (order: 2, window: 15 pt), SNV, DT	126	54	7	0.51	0.75	0.56	0.69	0.51	0.72
3	SNV, 1st der (order: 2, window: 15 pt)	126	54	5	0.66	0.58	0.7	0.54	0.58	0.65
4	2nd der (order: 3, window: 15 pt)	126	54	7	0.55	0.71	0.58	0.68	0.53	0.71
5	OSC, 2nd der (order: 3, window: 15 pt)	126	54	6	0.55	0.71	0.6	0.66	0.49	0.75
6	SNV, DT, 2nd der (order: 3, window: 15 pt)	126	54	6	0.58	0.68	0.62	0.64	0.51	0.73
7	**None**	**126**	**54**	**8**	**0.55**	**0.7**	**0.61**	**0.64**	**0.49**	**0.75**
CV	venetian blinds with 14 splits and 5 samples per split									

Triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R2$_C$	RMSECV	R2$_{CV}$	RMSEP	R2$_P$
1	SNV, DT	124	53	3	0.46	0.82	0.52	0.78	0.48	0.8
2	1st der (order: 2, window: 15 pt), SNV, DT	124	53	7	0.39	0.88	0.44	0.84	0.41	0.85
3	**SNV, 1st der (order: 2, window: 15 pt)**	**124**	**53**	**8**	**0.37**	**0.89**	**0.43**	**0.85**	**0.36**	**0.88**
4	2nd der (order: 3, window: 15 pt)	124	53	4	0.43	0.85	0.45	0.83	0.41	0.86
5	OSC, 2nd der (order: 3, window: 15 pt)	124	53	5	0.41	0.86	0.44	0.83	0.42	0.85
6	SNV, DT, 2nd der (order: 3, window: 15 pt)	124	53	4	0.4	0.86	0.45	0.83	0.38	0.88
7	None	124	53	5	0.43	0.85	0.47	0.81	0.37	0.87
CV	venetian blinds with 14 splits and 5 samples per split									

Combined data set

#	Preprocessing	XT	XV	LV	RMSEC	R2$_C$	RMSECV	R2$_{CV}$	RMSEP	R2$_P$
1	SNV, DT	249	107	13	0.43	0.84	0.48	0.79	0.49	0.81
2	1st der (order: 2, window: 15 pt), SNV, DT	249	107	12	0.43	0.83	0.49	0.79	0.5	0.81
3	SNV, 1st der (order: 2, window: 15 pt)	249	107	14	0.41	0.85	0.49	0.79	0.51	0.8
4	2nd der (order: 3, window: 15 pt)	249	107	12	0.42	0.84	0.47	0.8	0.5	0.81
5	OSC, 2nd der (order: 3, window: 15 pt,	249	107	14	0.41	0.85	0.47	0.8	0.5	0.82
6	**SNV, DT, 2nd der (order: 3, window: 15 pt)**	**249**	**107**	**12**	**0.42**	**0.84**	**0.48**	**0.79**	**0.49**	**0.82**
7	None	249	107	10	0.47	0.8	0.51	0.77	0.57	0.75
CV	venetian blinds with 14 splits and 5 samples per split									

* Standard normal variate (SNV), Detrend (DT), orthogonal signal correction (OSC)

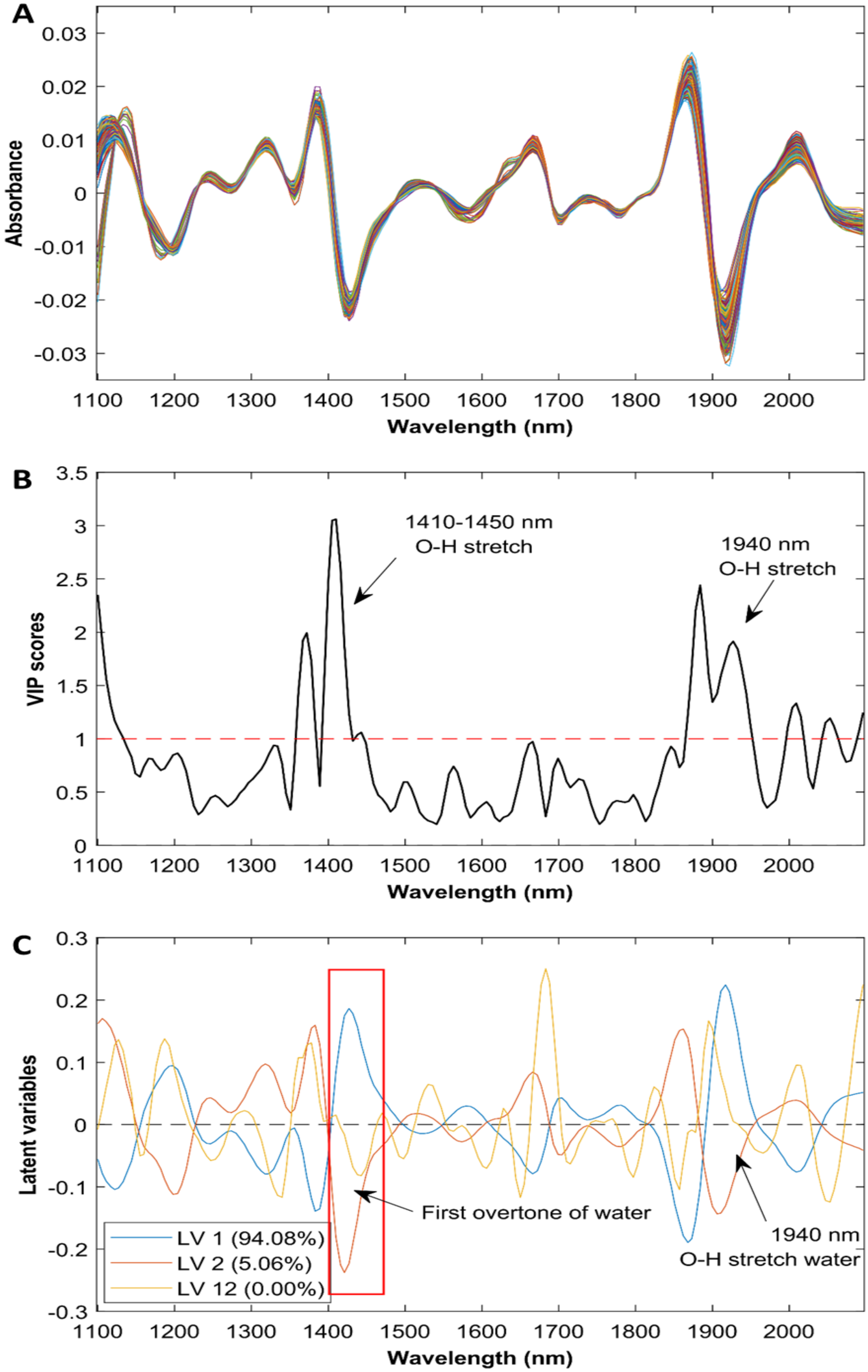

Figure 4.7. (a) Average pre-processed spectra of the combined data set for moisture content prediction model (Table 4.3), (b) variable importance in projection and (c) latent variables for LV 1, LV 2 and LV 11.

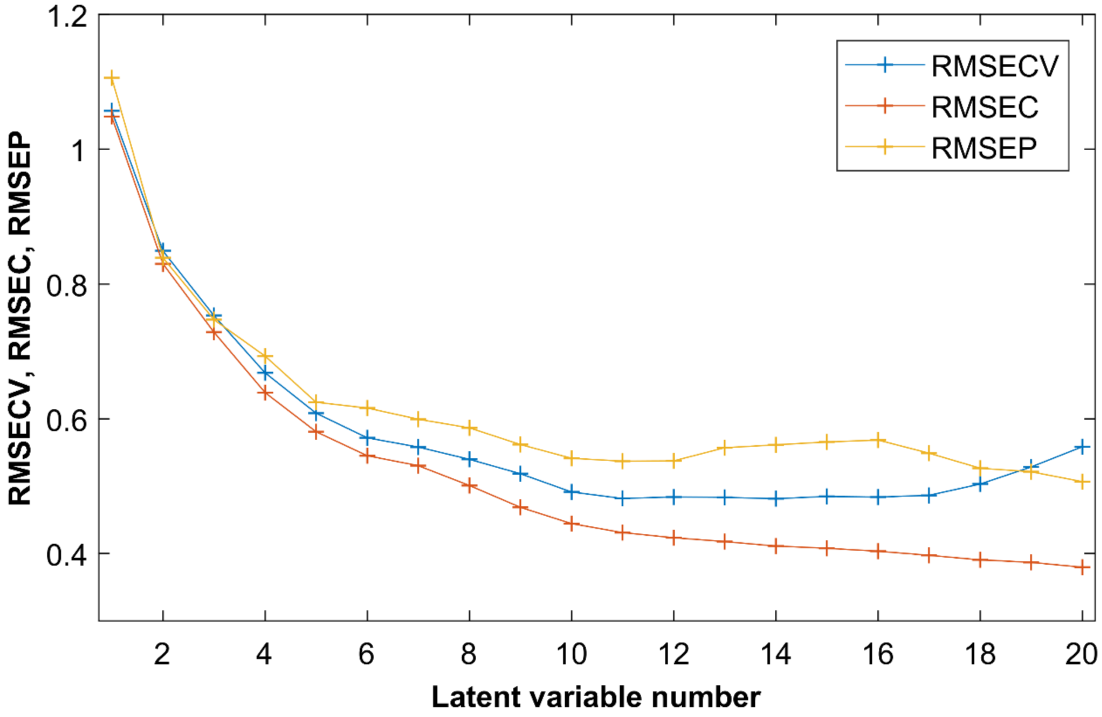

Figure 4.8 Latent variables vs. standard error of cross-validation, -calibration and -prediction error for the combined data set moisture prediction model 6 (Table 4.3).

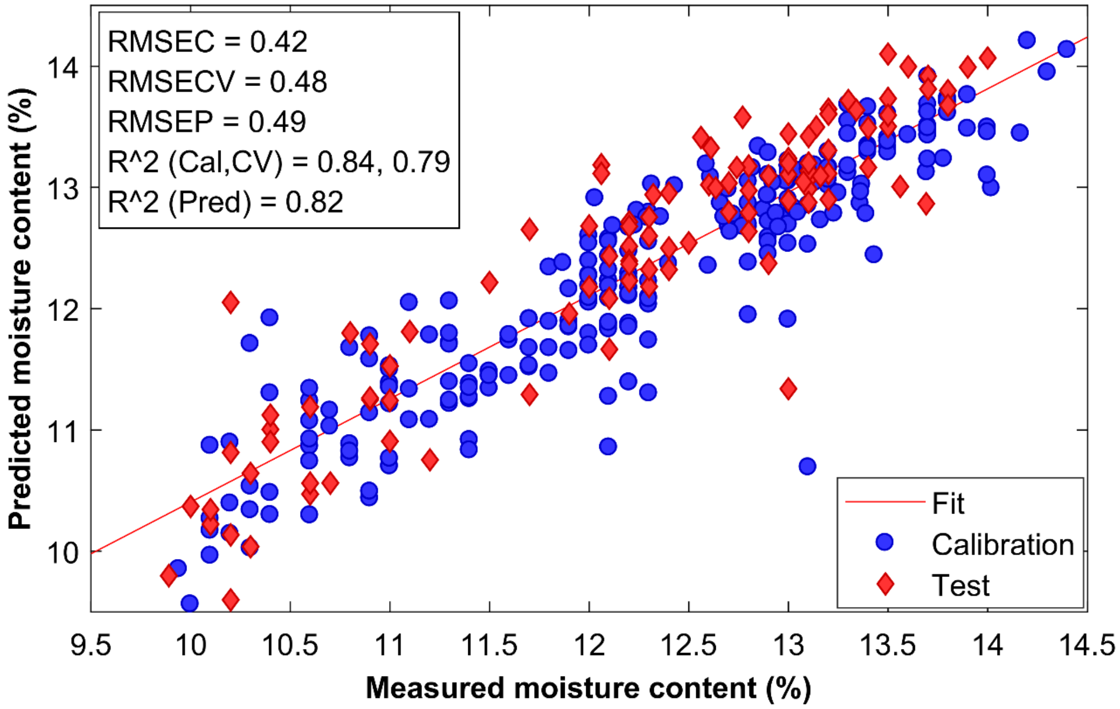

Figure 4.9 Measured vs. predicted moisture content for PLS regression model 6, with SNV DT and 2nd derivative pre-treatment for the combined wheat and triticale data set using 12 LV’s and 249 samples in the training set and 107 in the validation set.

4.2.3 Kernel hardness

Kernel hardness distribution histograms for wheat, triticale and the combined data sets are shown in Figure 4.10. Wheat average kernel hardness (64.87) was higher than that of triticale (54.24) and the triticale data set had a lower maximum (75.55) than wheat (84.96). The wheat data set had a higher standard deviation (SD) than the triticale data set and the combined data s*et al*so had a higher SD (Fig. 4.2.7)

The best PLS regression models for the bulk wheat data set (Table 4.4), spectrally pre-treated with 2nd derivative, resulted in an RMSEP of 5.56 with a R^2_P of 0.55. The triticale data set (Table 4.4) with no spectral pre-treatment resulted in a RMSEP of 4.71 with a R^2_P of 0.23 and for the combined wheat and triticale data set (Table 4.4) spectrally pre-treated with SNV and 2nd derivative resulted in a RMSEP of 8.66 with a R^2_P of 0.56. A large hardness SD between kernels within the same sample could have contributed to the poor regression models.

The plot of pre-treated spectra (Fig. 4.11a) for the combined wheat and triticale data set, shows the protuberance assigned to the wheat samples described in the PC1 vs. PC2 plot (Fig. 4.2). The VIP scores plot (Fig. 4.11b) for the same data set indicates variables of importance around 1100-1200 (carbonyls and alkenes), the first overtone of water (1460 nm), the protein absorption region (1460-1570 nm), the hydrocarbon region (1600-1730) and the cellulose to carboxylic acid regions (1820-1920 nm) (Williams *et al.*, 2019). The LV plot (Fig 4.11c) for LV1, LV2 and LV7 highlights the variables that contribute the most weight for hardness prediction, corresponding to the VIP scores plot.

The RMSEC, RMSECV and RMSEP plot for the increased number of LV's (Fig. 4.12) (combined data set) indicates overfitting after 7 LV's as the difference between RMSEC and RMSECV increases The measured vs. predicted bulk kernel hardness plot (Fig. 4.13) shows a large vertical spread around the regression line of best fit, this negatively contributes to prediction accuracy as these data points are considered vertical outliers.

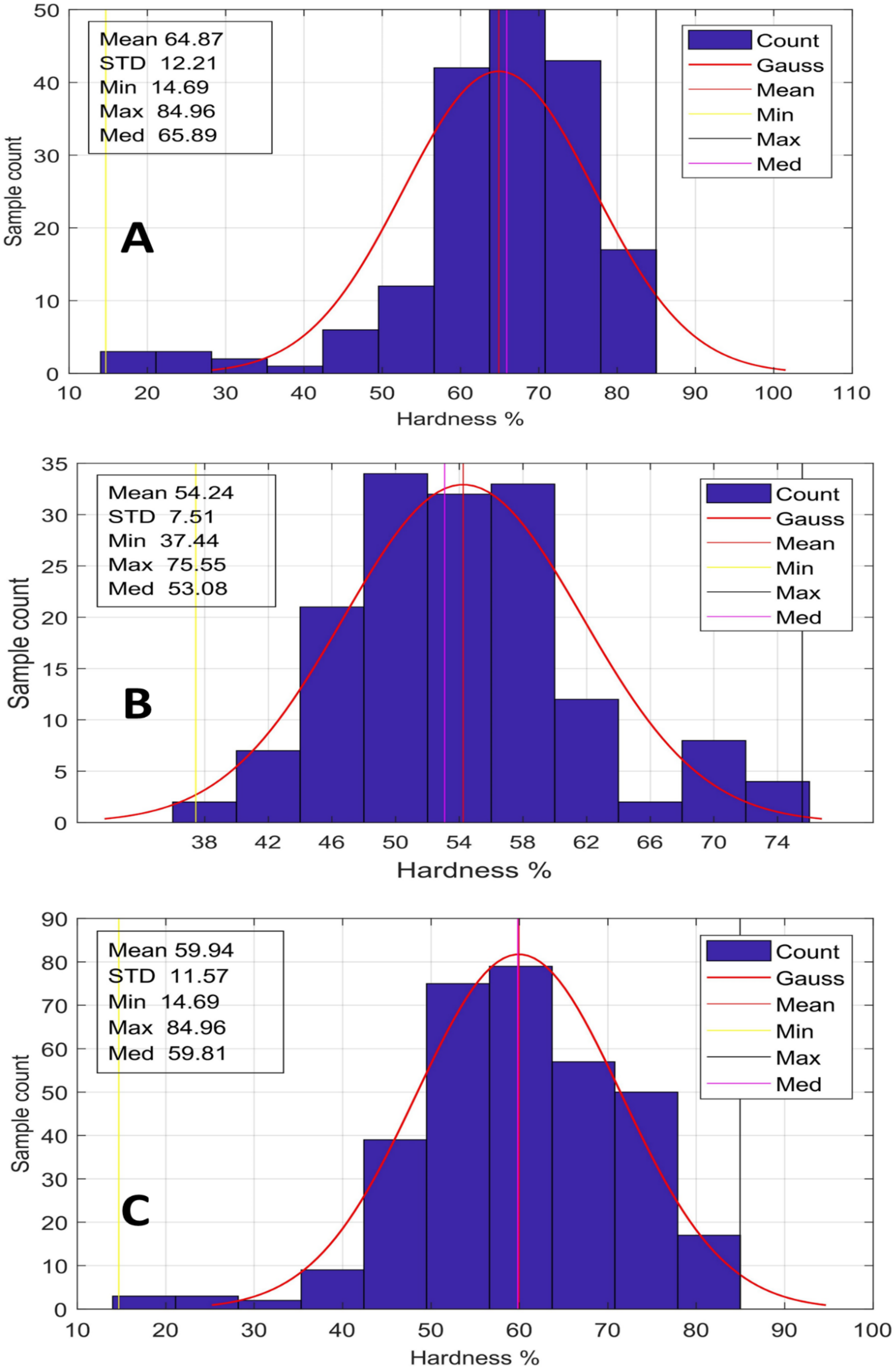

Figure 4.10 Histograms illustrating distribution of kernel hardness reference data for (a) wheat, (b) triticale and (c) the combined data sets.

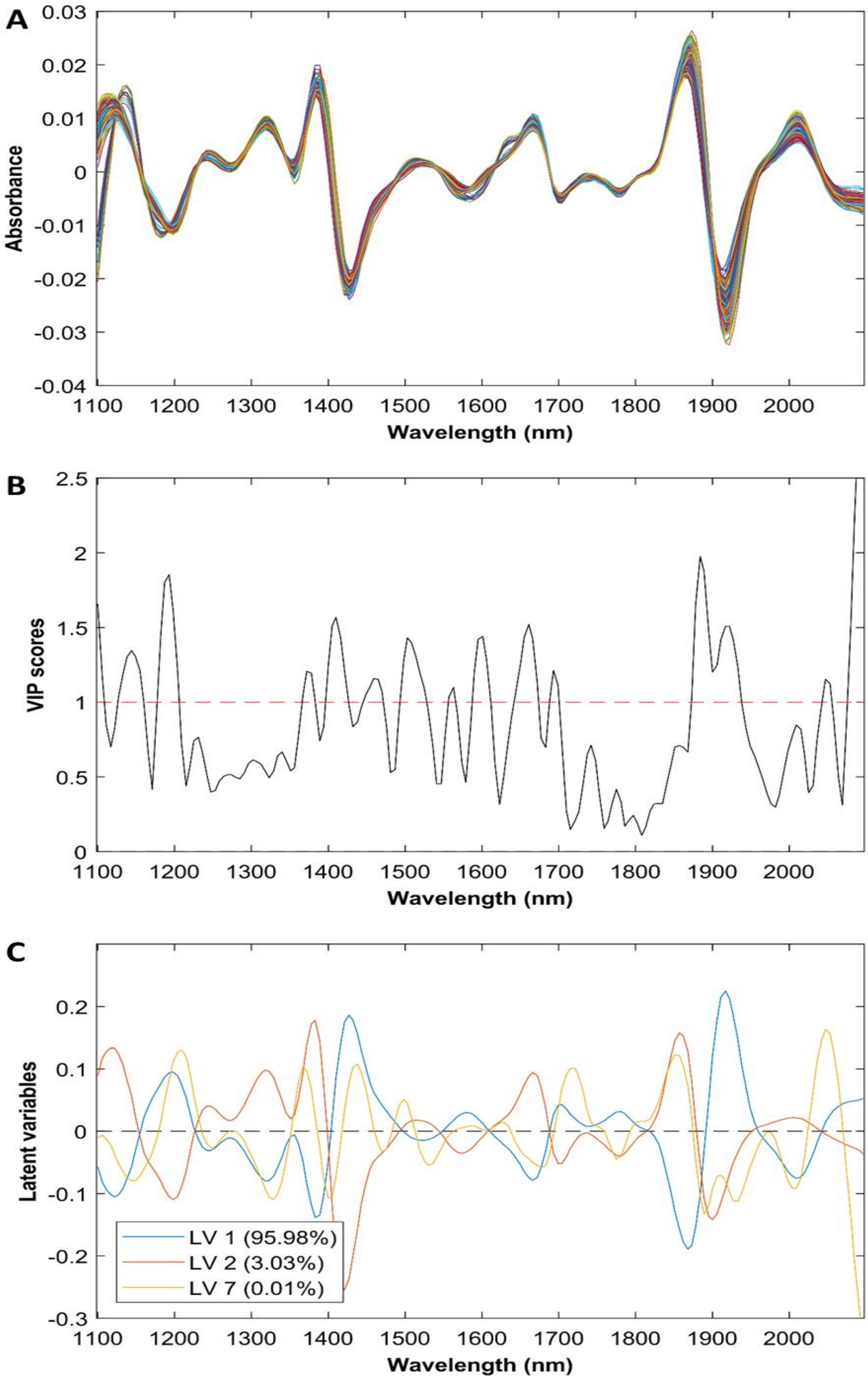

Figure 4.11. (a) Average pre-processed spectra of the combined data set for kernel hardness prediction model 1 (Table 4.4), (b) variable importance in projection and (c) LV's for (LV 1, LV 2 and LV 11).

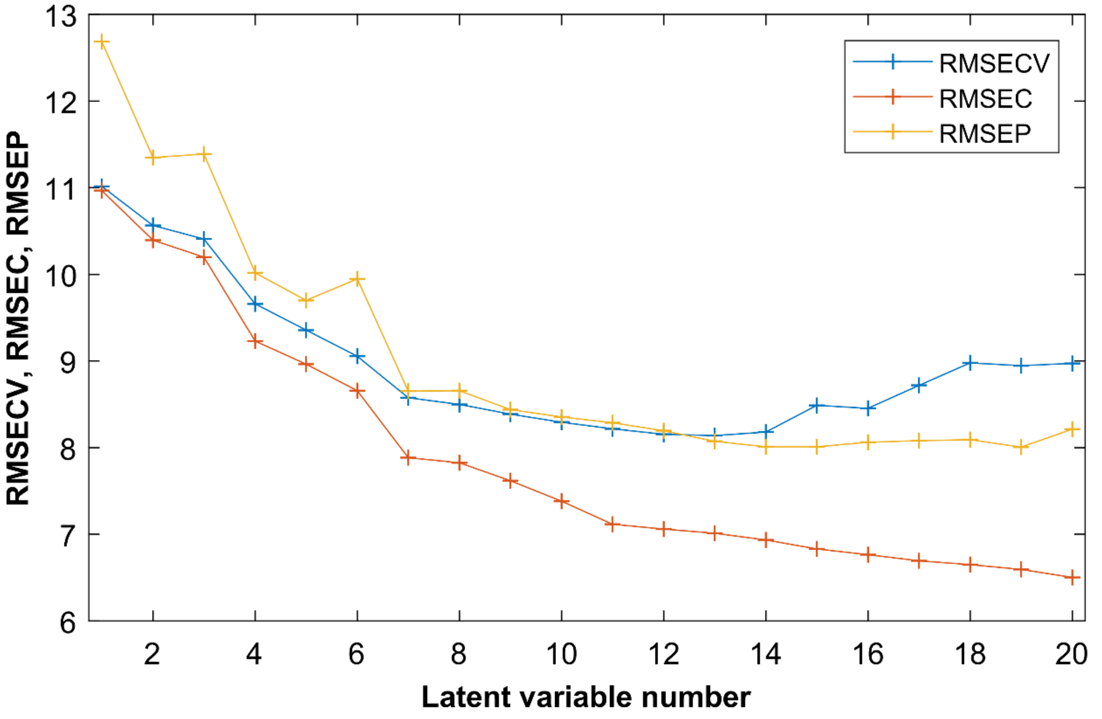

Figure 4.12 Latent variables vs standard error of cross-validation, -calibration and -prediction error for the combined data set hardness prediction model 1 (Table 4.4).

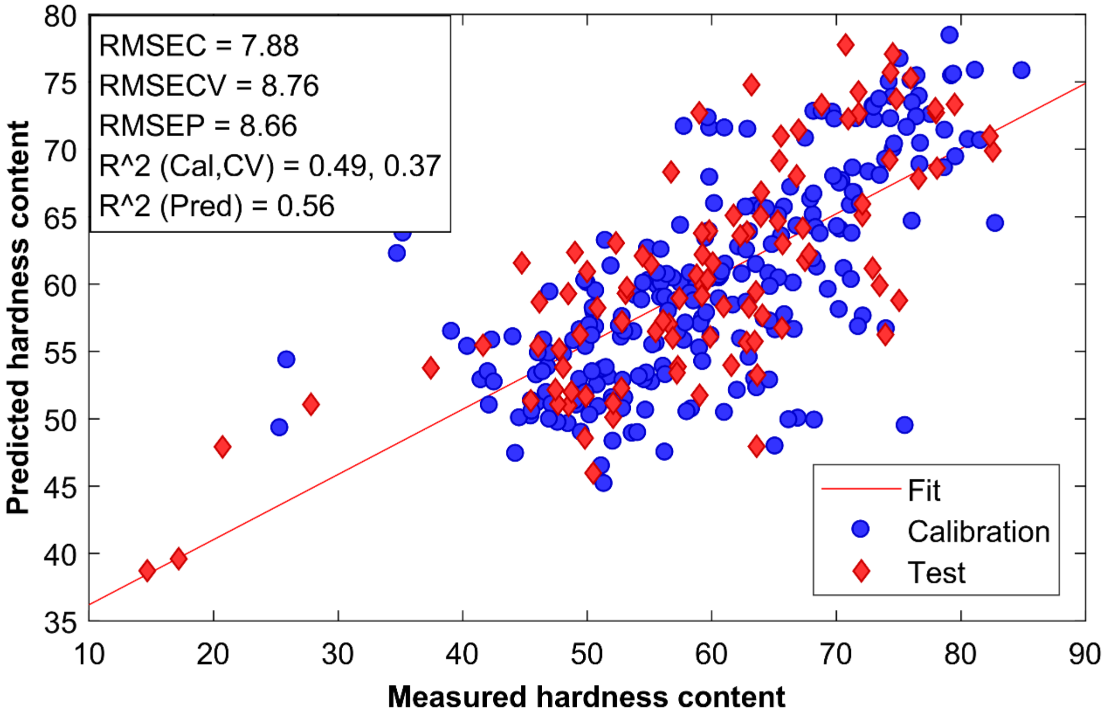

Figure 4.13 Measured vs. predicted kernel hardness values for the PLS regression model with SNV and 2nd derivative pre-treatment for the combined data set using 7 LV's and a calibration set of 234 samples and a validation set of 100.

Table 4.4 Calibration and validation statistics for predicted kernel hardness PLS regression models for bulk wheat, triticale and combined data sets using different pre-processing methods. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	SNV, 2nd der (order: 3, window: 15 pt)	116	50	6	4.95	0.63	5.30	0.58	10.22	0.53
2	SNV	116	50	4	5.33	0.57	5.61	0.53	6.07	0.48
3	SNV, DT	116	50	2	6.04	0.45	6.20	0.43	6.61	0.40
4	MC, SNV, DT	116	50	2	5.82	0.49	6.07	0.45	6.46	0.40
5	OSC	116	50	3	5.55	0.54	5.85	0.49	6.00	0.50
6	**2nd der (order: 3, window: 15 pt)**	**116**	**50**	**6**	**5.01**	**0.62**	**5.46**	**0.56**	**5.56**	**0.55**
7	None	116	50	3	5.51	0.54	5.72	0.51	5.84	0.52
CV	Venetian blind with 14 splits and 5 samples per split									

Triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	SNV, 2nd der (order: 3, window: 15 pt)	100	41	3	4.49	0.42	4.78	0.35	9.07	0.10
2	SNV	100	41	3	4.68	0.37	4.91	0.31	4.99	0.17
3	SNV, DT	100	41	3	4.70	0.37	4.96	0.30	5.29	0.09
4	MC, SNV, DT	100	41	3	4.79	0.34	5.20	0.26	5.38	0.05
5	OSC	100	41	2	4.55	0.41	4.79	0.35	5.37	0.08
6	2nd der (order: 3, window: 15 pt)	100	41	4	4.41	0.44	4.58	0.40	5.18	0.11
7	**None**	**100**	**41**	**2**	**5.02**	**0.28**	**5.13**	**0.25**	**4.71**	**0.23**
CV	Venetian blinds with 14 splits and 5 samples per split									

Combined data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	**SNV, 2nd der (order: 3, window: 15 pt)**	**234**	**100**	**7**	**7.88**	**0.49**	**8.58**	**0.40**	**8.66**	**0.56**
2	SNV	234	100	7	8.64	0.38	9.02	0.33	9.57	0.48
3	SNV, DT	234	100	8	8.38	0.42	9.02	0.33	9.31	0.48
4	MC, SNV, DT	234	100	6	8.90	0.35	9.25	0.29	10.83	0.28
5	OSC	234	100	6	9.49	0.26	10.03	0.18	11.29	0.23
6	2nd der (order: 3, window: 15 pt)	234	100	5	8.87	0.35	9.24	0.30	9.75	0.44
7	None	234	100	4	9.68	0.22	9.96	0.18	10.65	0.33
CV	Venetian blind with 14 splits and 5 samples per split									

*Standard normal variate (SNV), Detrend (DT), Orthogonal signal correction (OSC), Mean centring (MC)

Bulk wheat, triticale and combined data set models achieved good results which were better and also comparable to that of work done by other authors (Delwiche and Hruschka 2000; Maghirang and Dowell 2003; Igne *et al*., 2007; Manley *et al*., 2013; Mahesh *et al*., 2014). The models proved that the method can be used as a rapid, non-destructive and selective method for quantification of protein and moisture content and also kernel hardness for the wheat, triticale and combined wheat and triticale data set.

The combination of a wheat and triticale data sets for one model has of yet not been shown using NIR-HSI or conventional NIR spectroscopy. The results for the combined data set are comparable to results obtained in this study for the wheat and also the triticale data sets. The combined data set results are also comparable to the work of other authors such as by Delwiche and Hruschka, (2000). And in some instances the results of this study outperformed the results of other studies, this could be due to instrument improvements and advancement in data processing techniques. Previous studies which are comparable to the work done in this study on wheat and triticale separately on a bulk whole grain basis using near-infrared NIR spectroscopy have been shown by (Delwiche and Hruschka 2000; Maghirang and Dowell 2003; Igne *et al*., 2007; Manley *et al*., 2013) and using NIR-HSI (Mahesh *et al*., 2014). The results of this study on bulk wheat and also triticale data sets indicated that overall model performance increased when compared to the work of other authors. Of interest to note is that the spectral pre-treatment techniques used for the models, i.e. SNV and 2nd derivative were mostly the same as used by other authors. This indicates that for less complex data sets model performance can easily be achieved with the conventional spectral pre-treatment methods.

4.3 Single kernel prediction models

Two methods were evaluated for removal of outliers from the single kernel (SK) wheat, triticale and combined data set models. The first method using PLS regression with the Straightforward Implementation of a statistically inspired Modification of the PLS method (SIMPLS) algorithm and focussed on manually selecting outliers based on their position on a scores plot (score distance vs standardised residual). The criteria for selection and removal was based on data points being vertical outliers or having bad leverage with a standardised residual of more than 2 percent protein, moisture or a hardness index above 2. Bad leverage points are data points that do not follow the pattern of the majority of the data and have a significant negative impact towards good regression values. The second method used a robust SIMPLS algorithm (RSIMPLS), this method proved useful because SIMPLS focusses on the cross-covariance between the response and regressors combined with linear least squares regression and the results are often affected by abnormal data points. The RSIMPLS method starts by applying a robust PCA (ROBPCA) on x- and y-variables from the data set. Robust estimates replace the empirical cross-covariance between X and Y and the empirical covariance matrix and systematically moves on to the SIMPLS algorithm as these robust estimates are made. The ROBPCA method is orthogonally equivariant in the multidimensional PC space and this subsequently means that orthogonal data transformation leave the scores unchanged and loadings transformed appropriately. The now robust estimates are used to remove data points which show bad estimation and consequently have bad leverage. The technique is described fully by Hubert and Vanden Branden 2003. It was necessary to remove outliers from the data sets as a large proportion of observations fell within the ranges of having bad leverage or being vertical outliers.

4.3.1 Single kernel protein

Model results for robust-PLS and manual outlier removal for protein content prediction are shown in Table 4.5 and 4.6. Overall model prediction accuracies were better when the outliers were removed manually compared to using the robust-PLS method on the wheat, triticale and combined SK spectra data sets.

Multivariate spectral filtering techniques such as orthogonal signal correction and generalised least squares (OSC and GLS) proved useful in models obtained using the robust-PLS method. Models using GLS as spectral pre-treatment required less LV’s compared to conventional methods. Less LV’s are important for model simplification and improved computation performance. GLS down weighs sources of variance by correlating data that have similar reference values and removing data which does not fit the correlation.

Applying robust-PLS to the SK wheat protein data set and by spectral pre-treatment with GLS an RMSEP of 0.62 with an R^2_P of 0.66 was obtained (Table 4.5). The SK wheat protein content prediction model for the manual outlier removal method (pre-treated with SNV) resulted in an RMSEP of 0.37% with an R^2_P of 0.84 (Table 4.6). Spectra pre-treated with GLS resulted in less LV's (10) used and an RMSEP of 0.38% and R^2_P of 0.83. Less LV's show a decrease in model complexity, making for a robust model with a decrease in computation time.

Table 4.5 Calibration and validation statistics for predicted protein content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and the robust-PLS outlier removal method. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	**GLS (Y Gradient, alpha 0.0002)**	**3856**	**1562**	**11**	**0.56**	**0.65**	**0.60**	**0.61**	**0.62**	**0.66**
2	SNV, DT	3856	1562	18	0.60	0.60	0.61	0.58	0.63	0.64
3	SNV	3856	1562	19	0.60	0.69	0.61	0.59	0.63	0.64
4	DT	3856	1562	18	0.59	0.61	0.61	0.59	0.62	0.65
5	OSC	3856	1562	17	0.61	0.59	0.62	0.58	0.63	0.64
6	2nd der (order: 3, window: 15 pt)	3856	1562	16	0.60	0.60	0.62	0.58	0.63	0.64
7	None	3856	1562	16	0.62	0.58	0.63	0.56	0.65	0.62
CV	Venetian blinds with 20 splits and 5 samples per split									

Triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (Y Gradient, alpha 0.0002)	2838	1376	14	0.49	0.76	0.54	0.71	0.62	0.69
2	SNV, DT	2838	1376	16	0.56	0.69	0.57	0.67	0.65	0.66
3	SNV	2838	1376	17	0.56	0.68	0.57	0.67	0.63	0.65
4	DT	2838	1376	17	0.54	0.71	0.55	0.69	0.62	0.66
5	**OSC**	**2838**	**1376**	**18**	**0.54**	**0.71**	**0.55**	**0.70**	**0.61**	**0.67**
6	2nd der (order: 3, window: 15 pt)	2838	1376	18	0.55	0.70	0.56	0.68	0.64	0.64
7	None	2838	1376	18	0.54	0.70	0.56	0.70	0.61	0.67
CV	Venetian blind with 20 splits and 5 samples per split									

Combined data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	**GLS (Y Gradient, alpha 0.0002)**	**7672**	**3540**	**11**	**0.74**	**0.54**	**0.77**	**0.51**	**0.82**	**0.49**
2	SNV, DT	7672	3540	18	0.75	0.52	0.76	0.51	0.82	0.49
3	SNV	7672	3540	16	0.76	0.51	0.77	0.50	0.83	0.48
4	DT	7672	3540	16	0.76	0.51	0.77	0.50	0.83	0.47
5	OSC	7672	3540	16	0.77	0.50	0.78	0.49	0.84	0.47
6	2nd der (order: 3, window: 15 pt)	7672	3540	18	0.76	0.51	0.77	0.50	0.83	0.48
7	None	7672	3540	18	0.76	0.51	0.77	0.50	0.84	0.47
CV	Venetian blind with 20 splits and 5 samples per split									

Table 4.6 Calibration and validation statistics for predicted protein content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and removing outliers manually. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (Y Gradient, alpha 0.0002)	2883	1148	10	0.35	0.84	0.38	0.81	0.38	0.83
2	SNV, DT	2883	1148	17	0.36	0.83	0.37	0.82	0.37	0.84
3	**SNV**	**2883**	**1148**	**18**	**0.36**	**0.83**	**0.37**	**0.82**	**0.37**	**0.84**
4	DT	2883	1148	18	0.37	0.82	0.38	0.81	0.37	0.83
5	OSC	2883	1148	17	0.37	0.82	0.38	0.81	0.37	0.83
6	2nd der (order: 3, window: 15 pt)	2883	1148	18	0.37	0.82	0.38	0.81	0.38	0.83
7	None	2883	1148	18	0.37	0.82	0.38	0.83	0.37	0.83
CV	Venetian blind with 20 splits and 5 samples per split									

Triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (Y Gradient, alpha 0.0002)	2940	1208	13	0.51	0.79	0.45	0.75	0.44	0.77
2	**SNV, DT**	**2940**	**1208**	**18**	**0.42**	**0.77**	**0.43**	**0.76**	**0.44**	**0.78**
3	SNV	2940	1208	18	0.43	0.77	0.44	0.76	0.44	0.77
4	DT	2940	1208	18	0.43	0.76	0.44	0.75	0.45	0.76
5	OSC	2940	1208	18	0.44	0.76	0.45	0.75	0.45	0.76
6	2nd der (order: 3, window: 15 pt)	2940	1208	18	0.44	0.75	0.45	0.74	0.46	0.75
7	None	2940	1208	18	0.44	0.75	0.45	0.74	0.45	0.76
CV	Venetian blind with 20 splits and 5 samples per split									

Combined data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (generalized least squares) (Y Gradient, alpha 0.0002)	6459	2668	16	0.45	0.79	0.47	0.78	0.47	0.79
2	**SNV, DT**	**6459**	**2668**	**18**	**0.45**	**0.79**	**0.46**	**0.79**	**0.46**	**0.80**
3	SNV	6459	2668	18	0.46	0.79	0.46	0.79	0.46	0.80
4	DT	6459	2668	18	0.48	0.77	0.48	0.77	0.48	0.78
5	OSC	6459	2668	18	0.48	0.77	0.48	0.77	0.49	0.78
6	2nd der (order: 3, window: 15 pt)	6459	2668	18	0.48	0.77	0.48	0.77	0.49	0.78
7	None	6459	2668	15	0.51	0.74	0.51	0.74	0.52	0.75
CV	Venetian blind with 20 splits and 5 samples per split									

Generalised least squares (GLS), Standard normal variate (SNV), Detrend (DT) Orthogonal signal correction (OSC)

Figure 4.14a shows the SNV pre-treated spectra, VIP scores plot and LV plot for the wheat calibration set, on which manual removal of outliers was performed. Together with the VIP scores (Fig 4.14b) and the LV plot (Fig 4.14c), it indicates that there were no significant outliers within the spectra and that protein absorbance regions at the NH stretch (1430-1530 nm) and at the amide stretch (1960-2050 nm) carry weight to aid in wheat SK protein content prediction.

Figure 4.15 shows the RMSECV, RMSEC, and RMSEP values for increased number of LV's for the SK wheat data set (Table 4.6). Sixteen LV's were selected as model overfitting is not apparent and no large increase in prediction accuracy is noted after 16 LV's. Figure 4.16 shows the measured vs. predicted protein content plot for the SK wheat data set (Table 4.6). The plot shows a good spread around the range of measured reference data and no extreme vertical residuals are present.

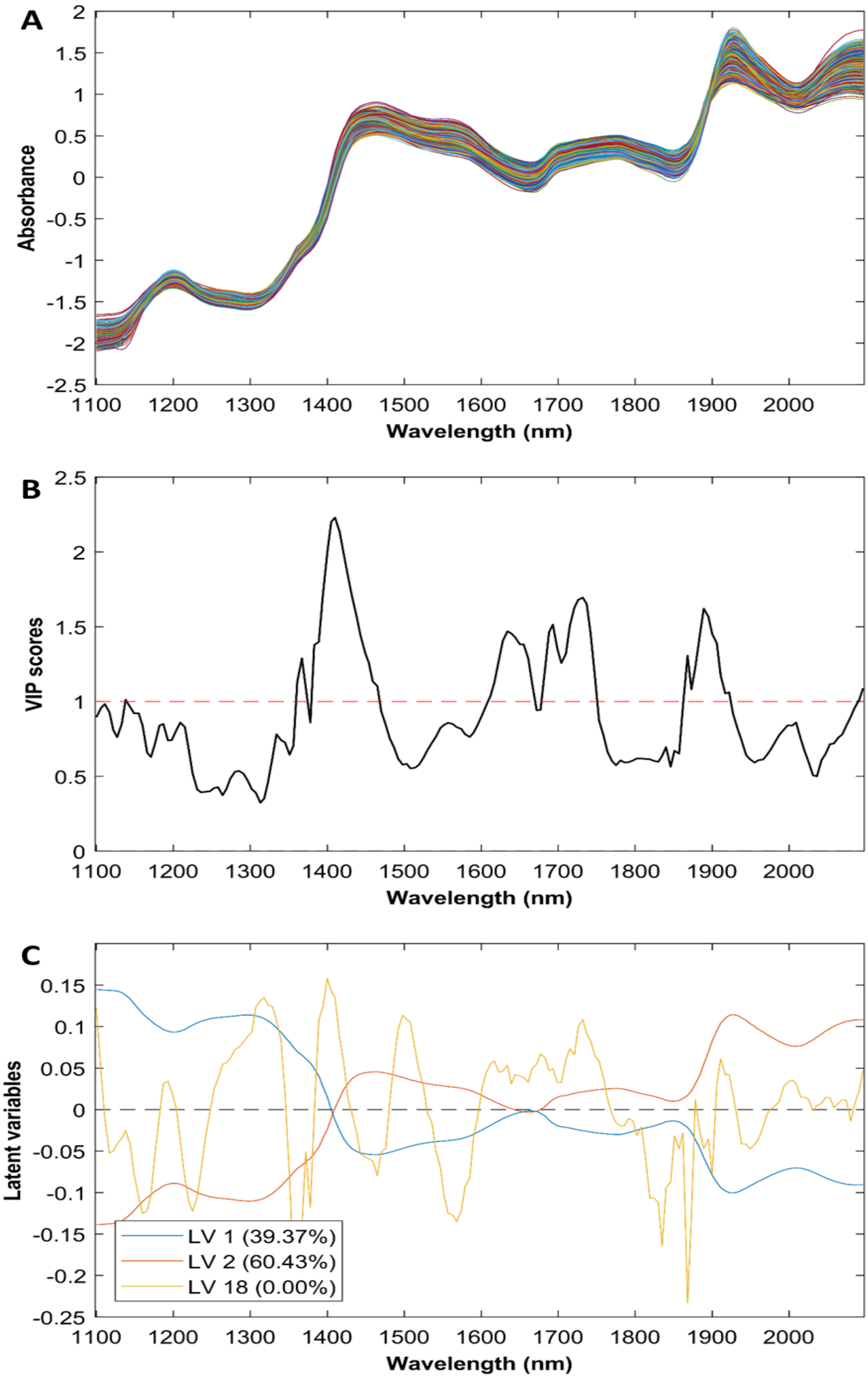

Figure 4.14 (a) Pre-processed SK spectra of the wheat data set for protein content prediction model 3 (Table 4.6), (b) variable importance in projection vs wavebands and (c) latent variables for LV 1, LV 2 and LV 18.

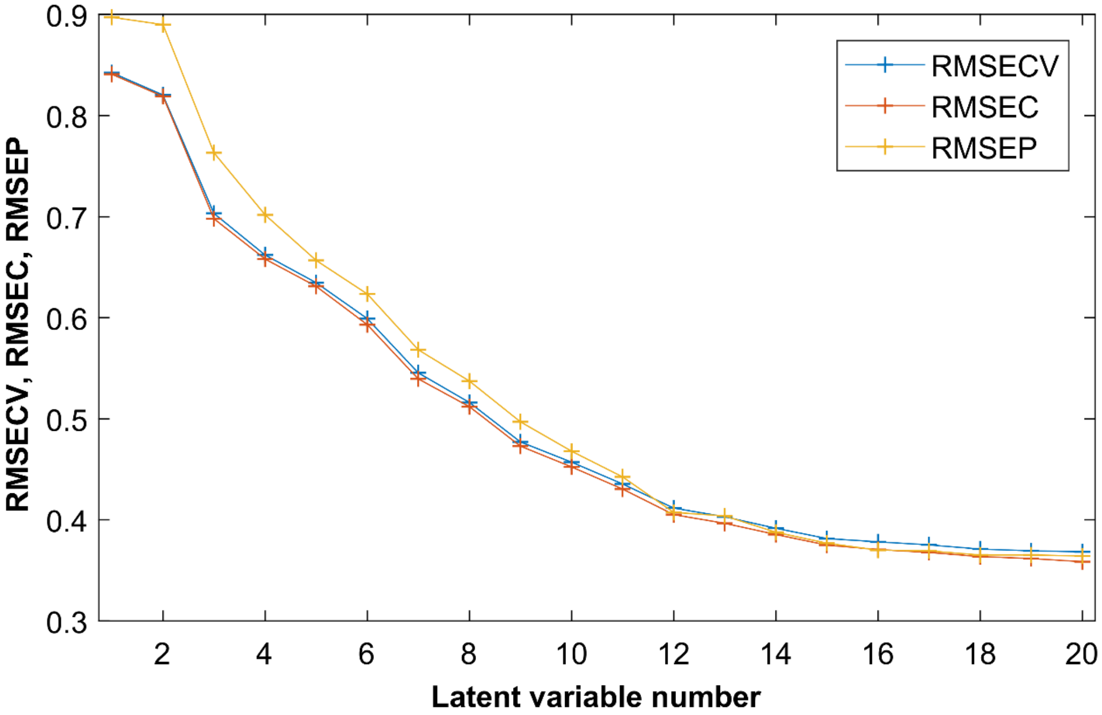

Figure 4.15. Latent variables vs standard error of cross-validation, -calibration and -prediction error for wheat protein prediction model 3 (Table 4.6).

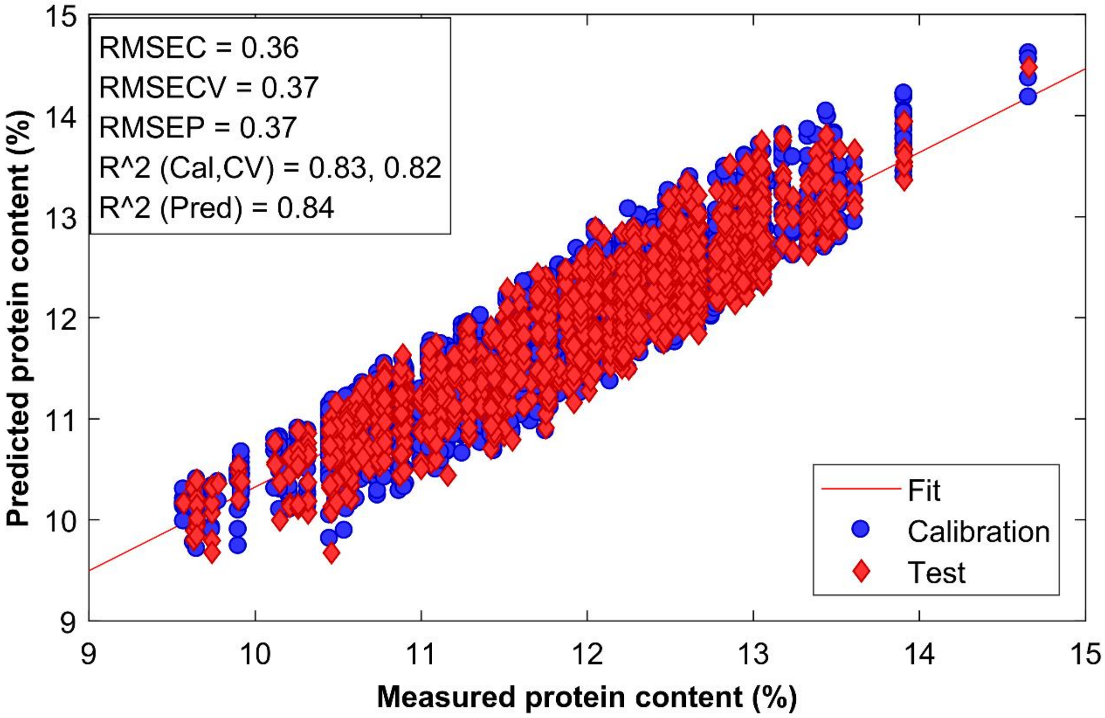

Figure 4.16 Measured vs. predicted protein content for PLS regression model 3 for the wheat data set pre-treated with SNV (Table 4.6) using 18 LV's and a calibration set of 2883 and validation set of 1148 single kernels.

Applying robust-PLS to the SK triticale protein content data set (Table 4.5), pre-treated with OSC (RMSEP of 0.61) and GLS (RMSEP of 0.62%; R^2_P of 0.69) resulted in less LV's used. Removing outliers manually (Table 4.6) with SNV and DT pre-treatments had the best model performance (RMSEP of 0.44% with a R^2_P of 0.78). The least LV's (13) were used with GLS (Table 4.6) and prediction results were a RMSEP of 0.44% and R^2_P of 0.77.

Figure 4.17a shows the SNV and DT pre-treated spectra, Figure 4.17b the VIP scores plot and Figure 4.17c the LV's plot for the SK triticale protein data set (model 2) of which outliers were removed manually (Table 4.6). The 1430-1530 nm region commonly associated with protein absorbance in NIR spectra, was not a significant variance of importance region for the triticale data set (Fig. 4.17b). The region at 1960-2050 nm was in turn indicated as an important variable region for protein content prediction. At 18 LV's in the latent variable plot (Fig. 4.17c) it was shown that the 1430-1530 nm wavelength area carries weight towards protein content prediction for the SK triticale data set.

Figure 4.18 shows the RMSEC, RMSECV and RMSEP for increased number of LV's for the SK triticale protein content data set applicable to model 2 (Table 4.6). With an increase in LV's a decrease in RMSECV, RMSEC and RMSEP was observed, with no overfitting being apparent before or after 18 LV's. The predicted vs measure SK triticale protein content plot (Fig. 4.19) shows no extreme vertical residuals which can be considered to be outliers, and shows a good distribution in line with model leverage.

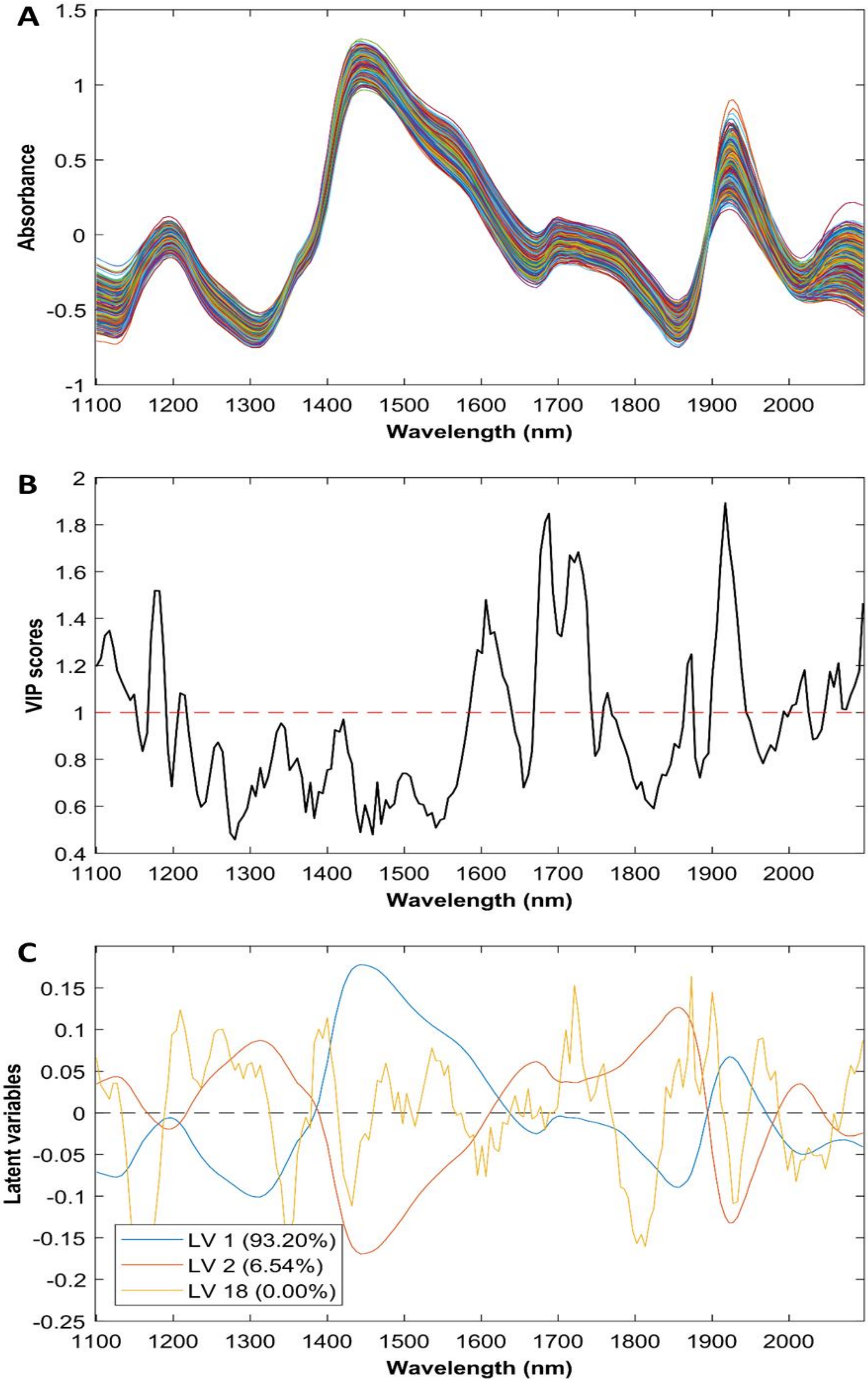

Figure 4.17 (a) Pre-processed spectra of the SK triticale data set for protein content prediction model 2 (Table 4.6), (b) variable importance in projection plot and (c) LV's plot for LV 1, LV 2 and LV 18.

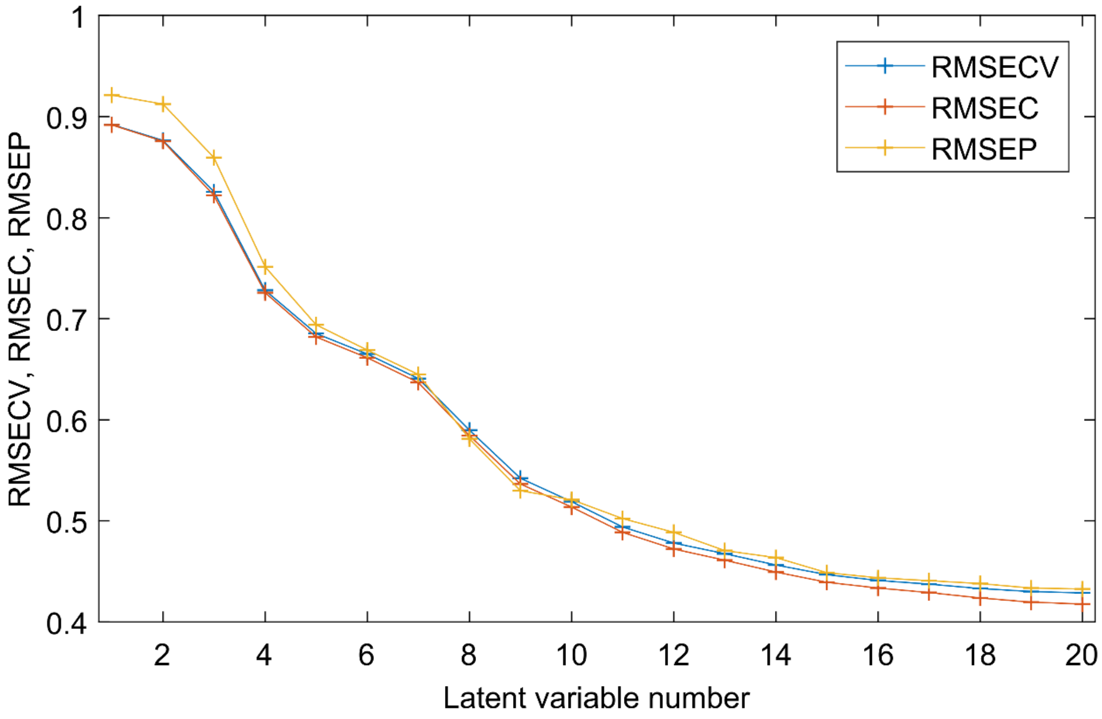

Figure 4.18 Latent variables vs standard error of cross-validation, -calibration and -prediction error for SK triticale protein content prediction model 2 (Table 4.6).

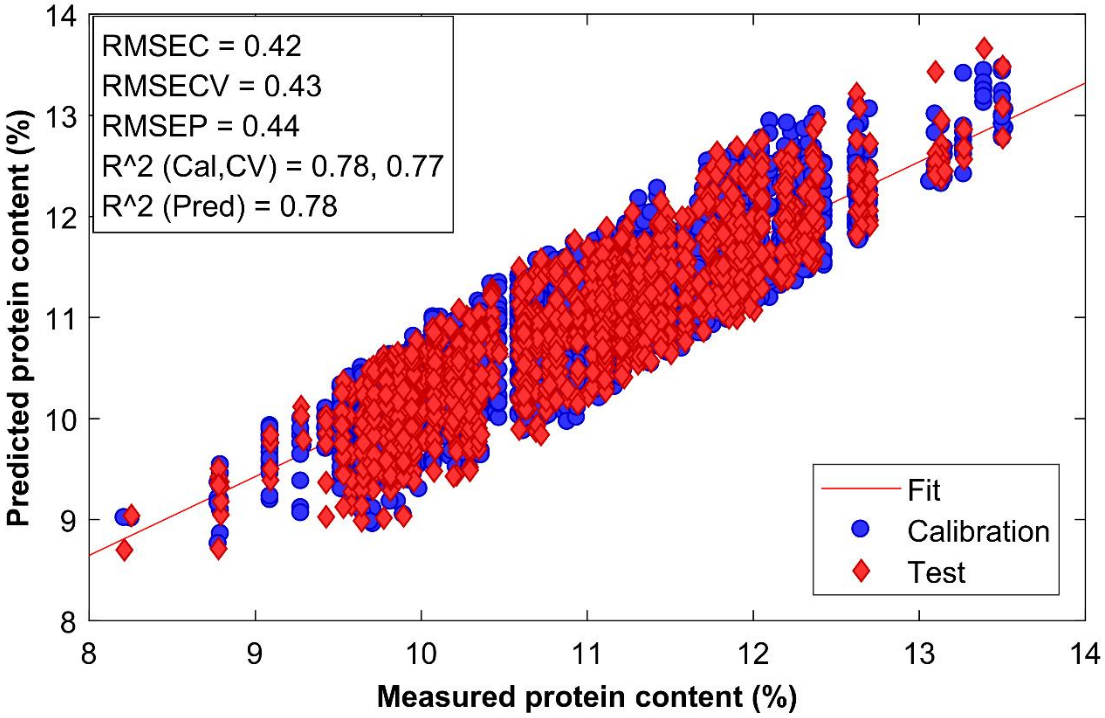

Figure 4.19 Measured vs. predicted protein content for PLS regression model 2 for SNV and DT pre-treated triticale SK data set (Table 4.6) with 18 LV’s and a calibration set of 2940 and a validation set of 1208 single kernels.

Applying robust-PLS and GLS to the combined SK data set for protein content determination (Table 4.5), resulted in a RMSEP of 0.82% with an R^2_P of 0.49. Removing outliers manually for SK protein content and using SNV and DT (Table 4.6) an RMSEP of 0.46% and R^2_P of 0.80 was obtained. Applying GLS to the SK combined (outliers manually removed) data set (Table 4.6) resulted in an RMSEP of 0.47 and an R^2_P of 0.79 with a reduced number of LV's (16) being used.

Figure 4.20a shows the SNV and DT pre-treated spectra for the combined SK data set (outliers manually removed), VIP scores plot (Fig. 4.20b) and LV's for LV1, LV2 and LV18 plot (Fig. 4.20c). The pre-treated (SNV and DT) spectra for the combined SK data set do not show the significant protrusion seen in Figure 4.2. Not having this protrusion present in the spectra gave an indication that multiplicative scatter effects were minimised by using the SK spectra. Figure 4.21 shows the VIP scores and LV's (LV1, LV2 and LV18) which indicates that the variables of importance for protein percentage prediction were also of significance, with the variables being clearly defined above a score of 1 at 1400 nm for the VIP scores plot and also at 18 LV's in the LV's plot.

The RMSECV, RMSEC and RMSEP values plotted against increasing number of LV's (Fig. 4.21) indicates an increase in model accuracy with an increase in LV's – no overfitting is observed with an increase in LV's. The predicted vs. measured combined SK data set plot (Fig. 4.22) indicates that reference values at the upper and lower limit of the calibration are under-represented. This indicates that these data points still fall within good leverage, but because they are under-represented they could also be responsible for skewing of prediction accuracy of the model.

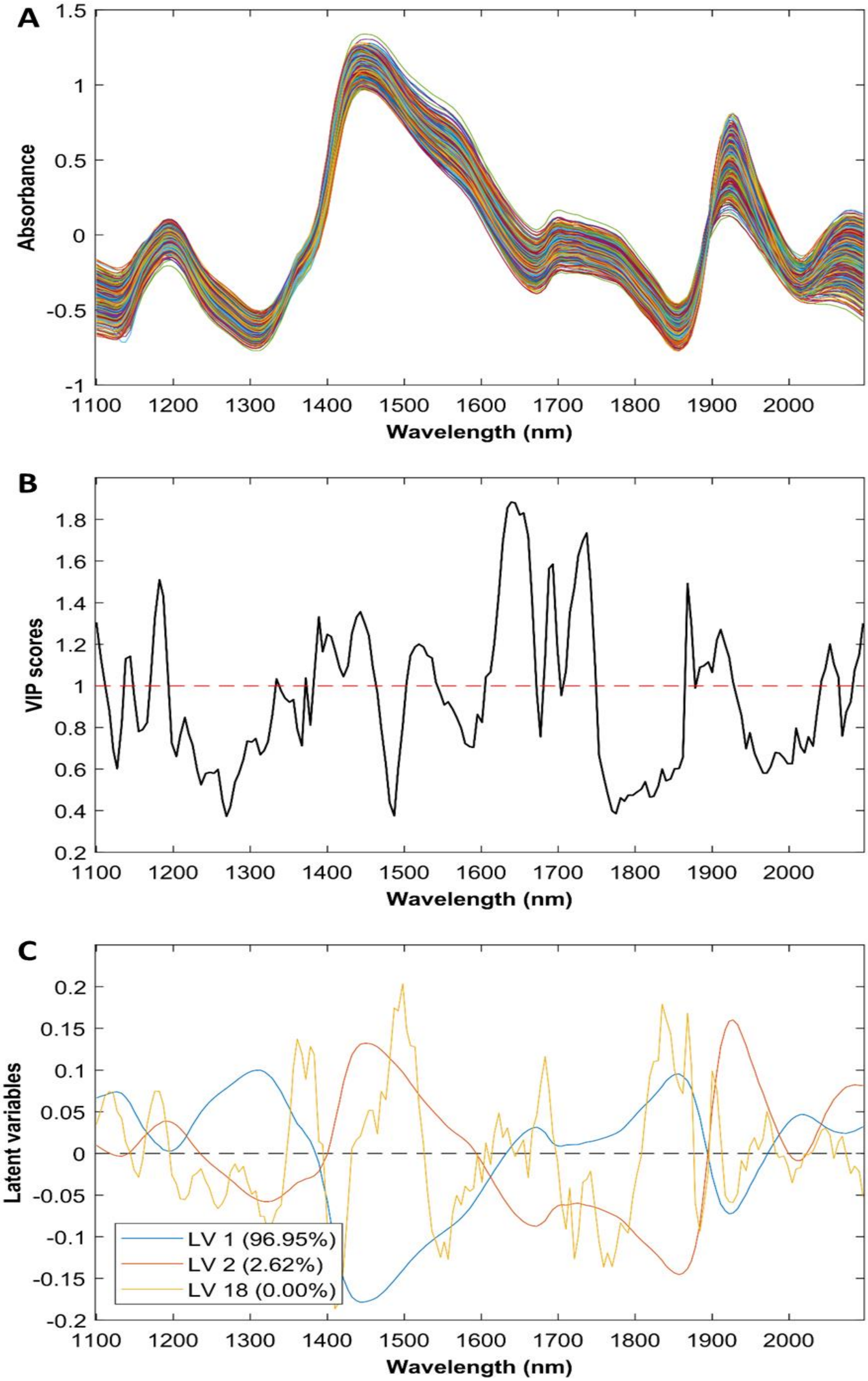

Figure 4.20. (a) Pre-processed spectra of the combined SK data set for protein content prediction model 2 (Table 4.6), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 18.

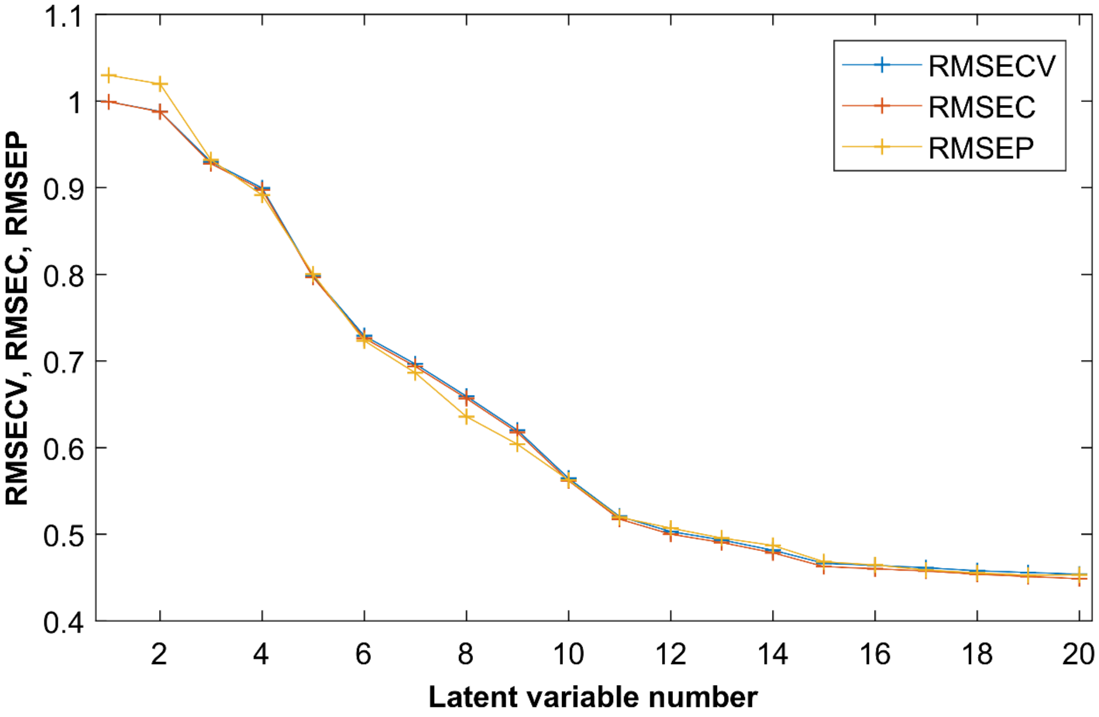

Figure 4.21 Latent variables vs standard error of cross-validation, -calibration and -prediction error for wheat and triticale protein prediction model 2 (Table 4.6).

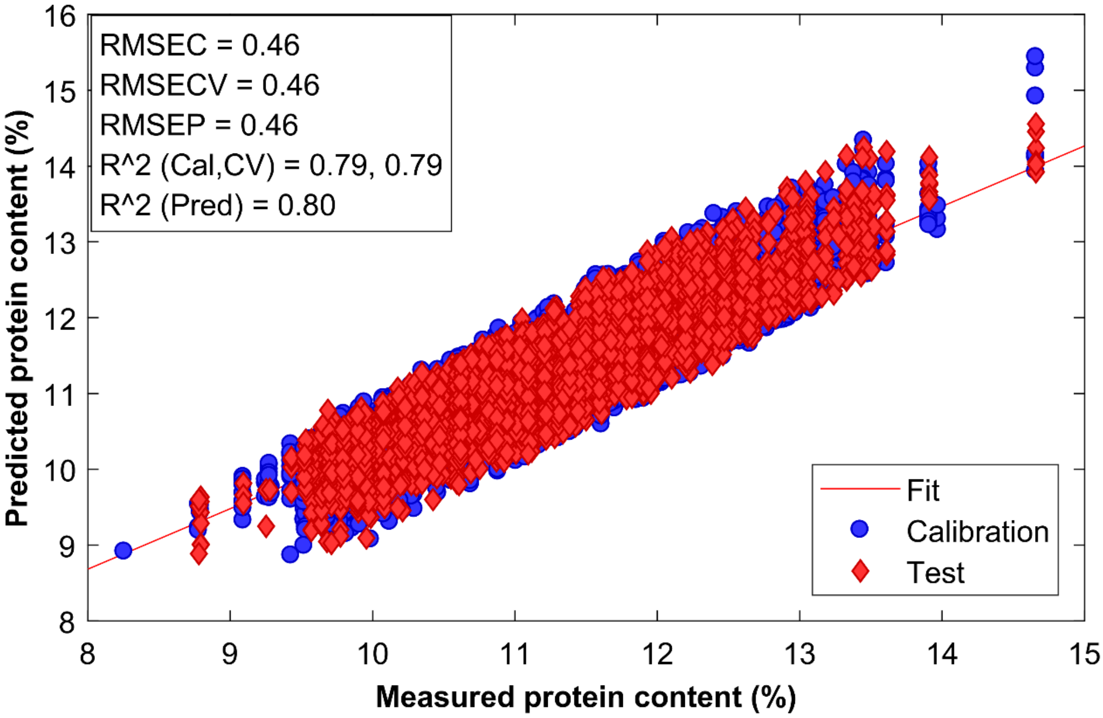

Figure 4.22 Measured vs. predicted protein content for PLS regression model with SNV and DT pre-treatment for the combined wheat and triticale SK data set (Table 4.6) using 18 LV’s and a calibration set of 6459 and 2669 single kernels.

4.4 Single kernel moisture content

Single kernel PLS regression moisture content models were calculated for wheat, triticale and for the combined wheat and triticale data set. Applying robust-PLS (RSIMPLS) and pre-treatment with SNV and DT resulted in an RMSEP of 0.50% and an R^2_P of 0.75 (Table 4.7) Applying GLS weighting to filter out spectral variance which is orthogonal to the reference data in the same data set (Table 4.7) an RMSEP of 0.50 and R^2_P of 0.75 were obtained with only 6 LV's. Outliers were removed manually from the data set and SNV and DT was applied to the spectra (Table 4.8) resulting in an RMSEP of 0.24% with an R^2_P of 0.93. When GLS was applied to the same data set (Table 4.8) an RMSEP of 0.28% and R^2_P of 0.91 using 6 LV's was obtained.

Figure 4.23 shows the SNV pre-treated spectra, the VIP scores and the LV's for LV1, LV2 and LV16. At the first overtone of water (1450 nm) significant absorbance is present in the pre-treated spectra. Figure 4.23b showing the VIP scores indicates that the first overtone of water at 1450 nm and at the OH combination band at 1940 nm are significant for moisture content prediction. Furthermore, Figure 4.23c showing LV's justifies that at LV 1, LV 2 and LV 16, the variables highlighted in Figure 4.23b (1450 and 1940 nm) are indeed significant.

Figure 4.24 shows a decrease in error (RMSECV, RMSEC and RMSEP) with increase in number of LV's. Model overfitting is not observed as an increase in number of LV's resulted in only a marginal increase in model accuracy. Figure 4.25 shows the predicted vs. measured SK wheat moisture content, the plot indicates no extreme vertical residuals and a good leverage of the predicted and measured data is observed.

Table 4.7 Calibration and validation statistics for predicted moisture content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and the robust-PLS outlier removal method. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (Y Gradient, alpha 0.0002)	4227	2106	6	0.39	0.84	0.41	0.82	0.50	0.75
2	**SNV, DT**	**4227**	**2106**	**16**	**0.39**	**0.83**	**0.40**	**0.84**	**0.50**	**0.75**
3	SNV	4227	2106	16	0.40	0.84	0.40	0.83	0.50	0.75
4	DT	4227	2106	16	0.44	0.80	0.45	0.79	0.55	0.71
5	OSC	4227	2106	16	0.44	0.80	0.45	0.79	0.54	0.72
6	2nd der (order: 3, window: 15 pt)	4227	2106	16	0.44	0.80	0.45	0.80	0.55	0.70
7	None	4227	2106	17	0.46	0.80	0.45	0.79	0.54	0.72
CV	Venetian blind with 20 splits and 5 samples per split									

Triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (Y Gradient, alpha 0.0002)	3996	2070	9	0.41	0.85	0.43	0.84	0.47	0.81
2	**SNV, DT**	**3996**	**2070**	**16**	**0.42**	**0.85**	**0.42**	**0.84**	**0.46**	**0.82**
3	SNV	3996	2070	13	0.43	0.84	0.43	0.83	0.47	0.81
4	DT	3996	2070	13	0.45	0.82	0.46	0.82	0.51	0.78
5	OSC	3996	2070	16	0.44	0.83	0.45	0.82	0.49	0.80
6	2nd der (order: 3, window: 15 pt)	3996	2070	16	0.45	0.82	0.46	0.81	0.51	9.78
7	None	3996	2070	14	0.45	0.82	0.45	0.82	0.50	0.79
CV	Venetian blind with 20 splits and 5 samples per split									

Wheat and triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (generalized least squares) (Y Gradient, alpha 0.0002)	8380	4200	9	0.46	0.81	0.47	0.80	0.53	0.78
2	**SNV, DT**	**8380**	**4200**	**16**	**0.47**	**0.80**	**0.57**	**0.80**	**0.51**	**0.79**
3	SNV	8380	4200	16	0.47	0.80	0.47	0.80	0.52	0.78
4	DT	8380	4200	16	0.49	0.78	0.49	0.78	0.55	0.76
5	OSC	8380	4200	16	0.49	0.78	0.49	0.78	0.56	0.75
6	2nd der (order: 3, window: 15 pt)	8380	4200	16	0.51	0.77	0.51	0.76	0.57	0.74
7	None	8380	4200	16	0.49	0.78	0.50	0.78	0.56	0.75
CV	Venetian blind with 20 splits and 5 samples per split									

Table 4.8 Calibration and validation statistics for predicted moisture content PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and removing outliers manually. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_c	RMSECV	R^2_{cv}	RMSEP	R^2_p
1	GLS (Y Gradient, alpha 0.0002)	3803	1630	6	0.27	0.92	0.28	0.91	0.28	0.91
2	SNV, DT	3803	1630	17	0.23	0.94	0.24	0.94	0.24	0.93
3	**SNV**	**3803**	**1630**	**16**	**0.23**	**0.94**	**0.24**	**0.94**	**0.24**	**0.94**
4	DT	3803	1630	16	0.30	0.90	0.30	0.90	0.31	0.90
5	OSC	3803	1630	16	0.29	0.91	0.30	0.90	0.30	0.90
6	2nd der (order: 3, window: 15 pt)	3803	1630	15	0.31	0.90	0.31	0.89	0.33	0.88
7	None	3803	1630	17	0.29	0.91	0.30	0.90	0.30	0.90
CV	Venetian blind with 20 splits and 5 samples per split									

Triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_c	RMSECV	R^2_{cv}	RMSEP	R^2_p
1	GLS (Y Gradient, alpha 0.0002)	3296	1325	8	0.27	0.93	0.27	0.92	0.29	0.92
2	SNV, DT	3296	1325	14	0.28	0.92	0.28	0.92	0.27	0.93
3	**SNV**	**3296**	**1325**	**14**	**0.28**	**0.92**	**0.28**	**0.92**	**0.27**	**0.93**
4	DT	3296	1325	14	0.31	0.91	0.31	0.91	0.31	0.91
5	OSC	3296	1325	17	0.29	0.92	0.30	0.91	0.30	0.92
6	2nd der (order: 3, window: 15 pt)	3296	1325	16	0.32	0.90	0.32	0.90	0.33	0.90
7	None	3296	1325	16	0.30	0.91	0.30	0.91	0.30	0.91
CV	Venetian blind with 20 splits and 5 samples per split									

Wheat and triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_c	RMSECV	R^2_{cv}	RMSEP	R^2_p
1	GLS (Y Gradient, alpha 0.0002)	6722	2894	8	0.27	0.93	0.28	0.92	0.29	0.92
2	SNV, DT	6722	2894	16	0.23	0.95	0.23	0.95	0.23	0.95
3	**SNV**	**6722**	**2894**	**17**	**0.23**	**0.95**	**0.23**	**0.95**	**0.23**	**0.95**
4	DT	6722	2894	17	0.29	0.92	0.29	0.92	0.30	0.92
5	OSC	6722	2894	17	0.28	0.92	0.28	0.92	0.29	0.92
6	2nd der (order: 3, window: 15 pt, tails: weighted)	6722	2894	15	0.32	0.90	0.32	0.90	0.33	0.90
7	None	6722	2894	15	0.29	0.91	0.30	0.91	0.31	0.91
CV	Venetian blind with 20 splits and 5 samples per split									

Generalised least squares (GLS), Standard normal variate (SNV), Detrend (DT), Orthogonal signal correction (OSC)

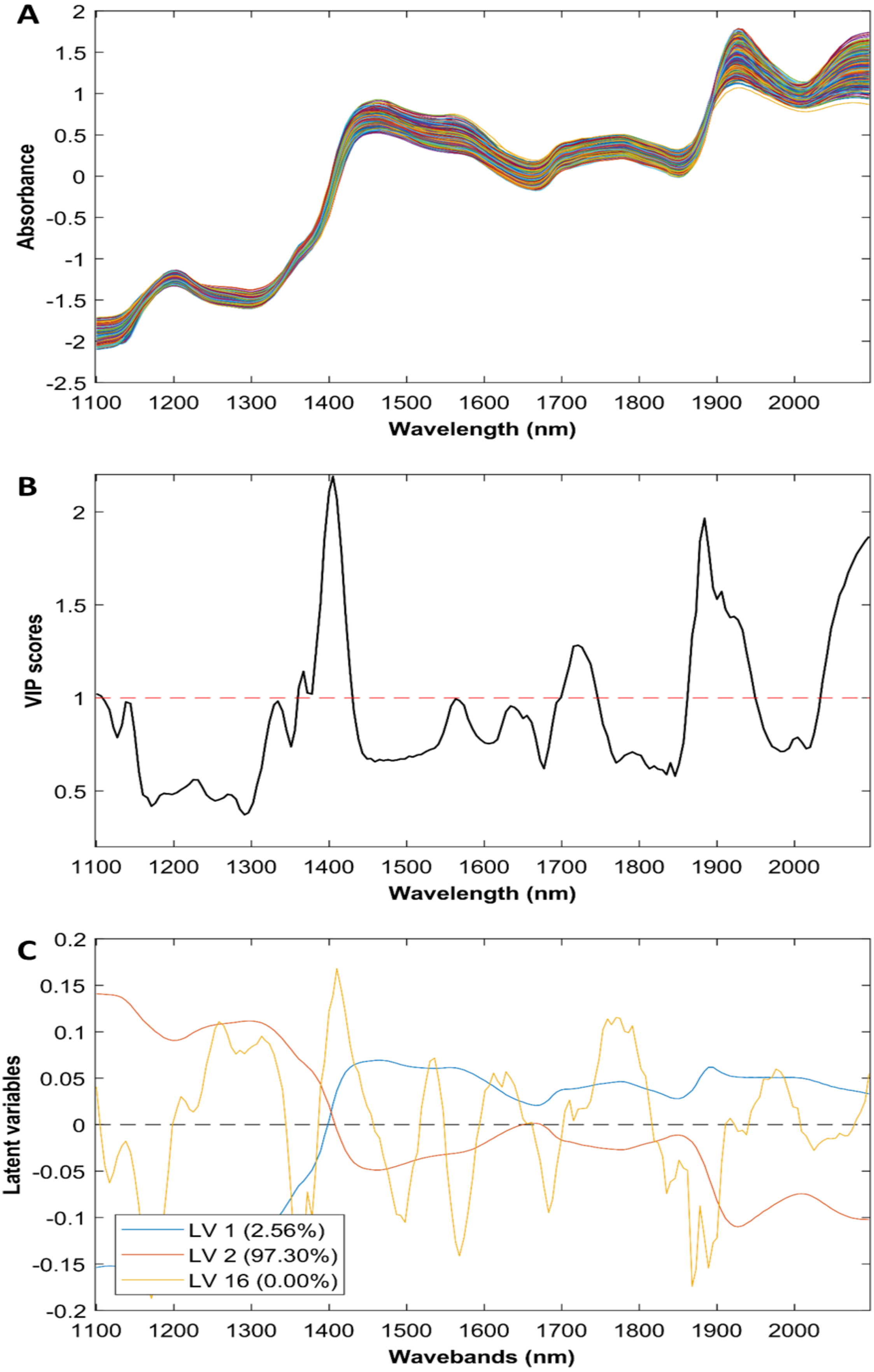

Figure 4.23 (a) Pre-processed spectra of the wheat SK data set for moisture content prediction model 3 (Table 4.8) wheat moisture content prediction, (b) variable importance in projection plot and (c) LV's for LV 1, LV 2 and LV.

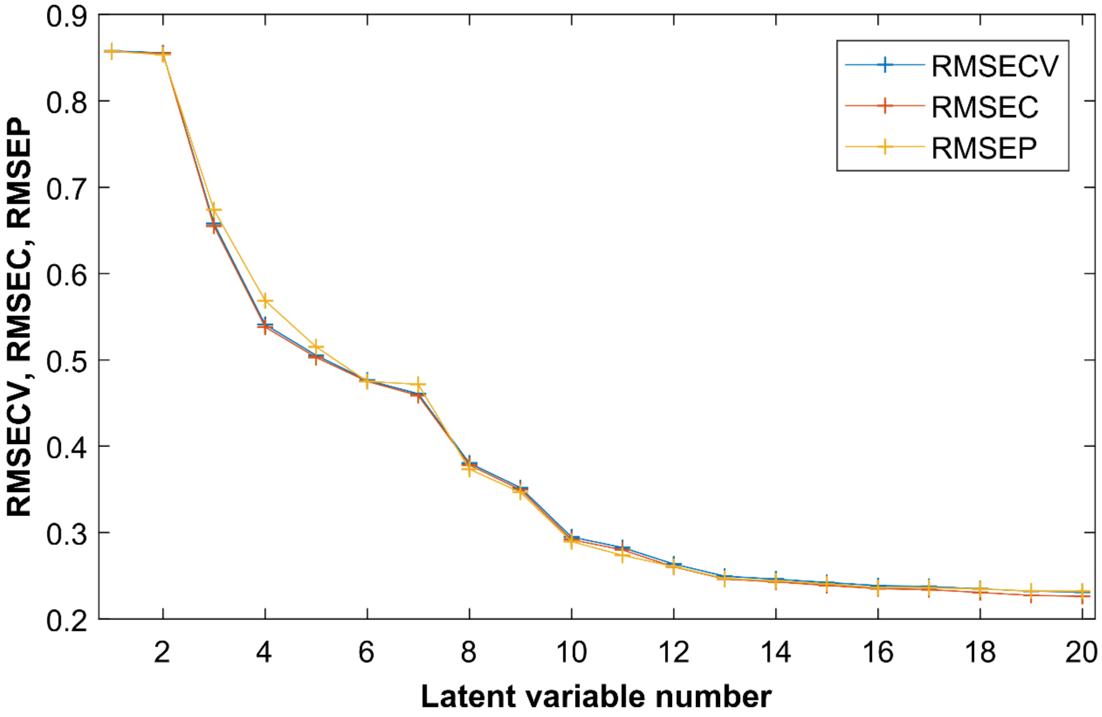

Figure 4.24 Latent variables vs standard error of cross-validation, -calibration and -prediction error for SK wheat moisture content prediction model 3 (Table 4.8).

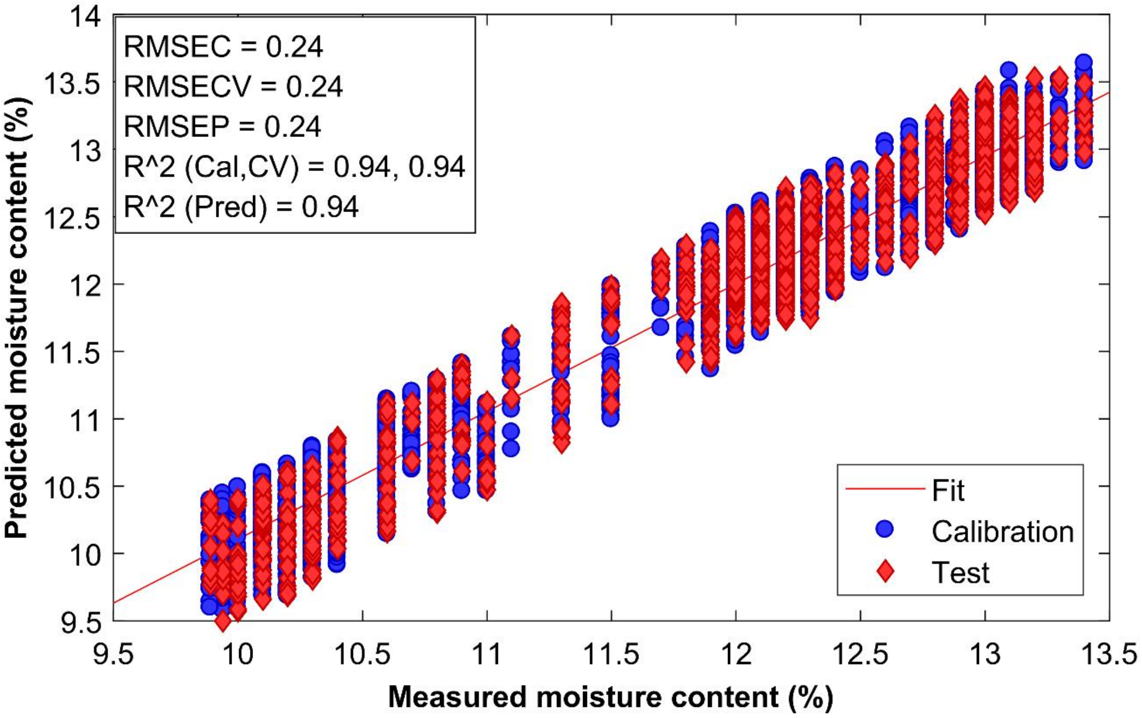

Figure 4.25 Measured vs. predicted moisture content for PLS regression model with SNV pre-treatment for the SK wheat data set model 3 (Table 4.8) using 16 LV's and a calibration set of 3803 and a validation set of 1630 SK's.

Triticale moisture content prediction using robust-PLS (SNV and DT) on the SK data set resulted in an RMSEP of 0.46% with an R^2_P of 0.82 (Table 4.7). Using GLS as spectral pre-treatment an RMSEP of 0.47% with an R^2_P of 0.81 using 9 LV's was obtained. By manually removing outliers from the triticale SK data set and by pre-treatment with SNV (Table 4.8) an RMSEP of 0.27% with an R^2_P of 0.93 was achieved. When GLS was used as spectral pre-treatment on the same SK triticale data set 8 LV's were used to obtain an RMSEP of 0.29 and an R^2_P of 0.92.

Figure 4.26a shows the SNV pre-treated SK triticale spectra, it indicates a good multiplicative scatter corrected set of SK spectra with no visual abnormalities. The VIP scores and LV's for LV1, LV2 and LV14 are shown in Figures 4.26b and 4.26c. The plots indicate that the spectral absorbance regions assigned to the first overtone of water (1450 nm) and the OH combination bands around 1940 nm carry significant weight towards moisture content prediction for the triticale data set.

The RMSECV, RMSEC and RMSEP plot for increasing number of LV's (Fig. 4.27) shows no overfitting and a decrease in error with increase in LV's. The predicted vs. measured values for the SNV triticale moisture content data set (Table 4.8) is shown in Figure 4.28. The plot indicates a good spread of predicted and measured data points with good leverage and no vertical outliers are indicated after manual outlier removal.

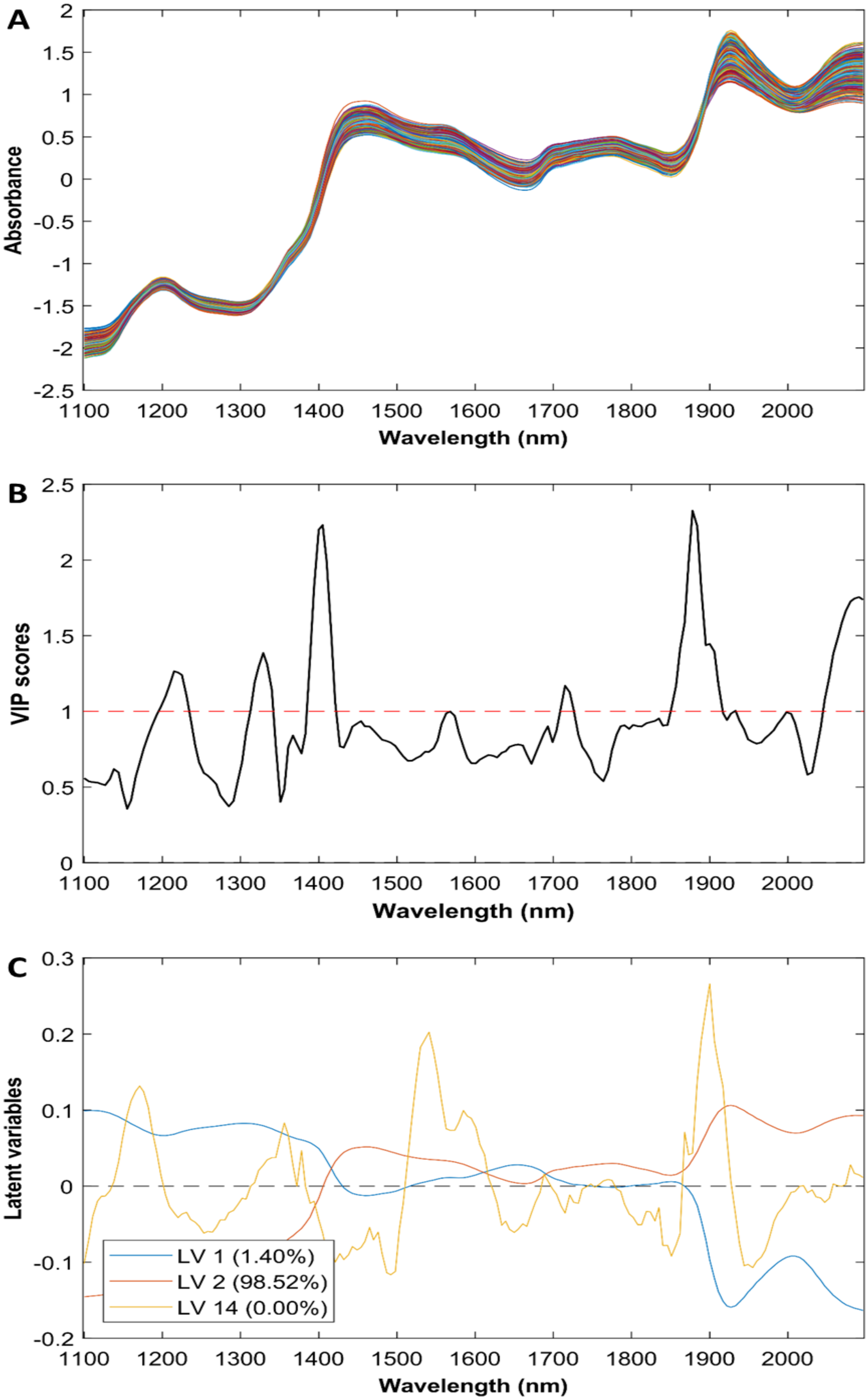

Figure 4.26 (a) Pre-processed spectra of the triticale SK data set for moisture content prediction for model 3 (Table 4.8), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 14.

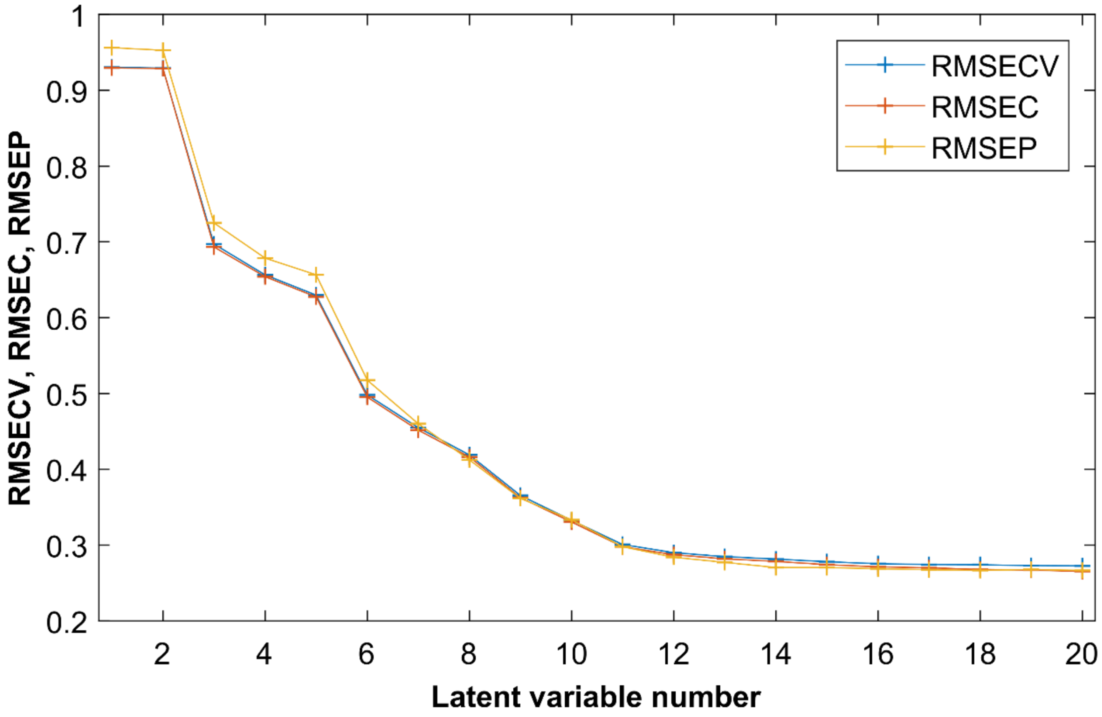

Figure 4.27 Latent variables vs standard error of cross validation, -calibration and -prediction for SK triticale moisture content prediction model 3 (Table 4.8).

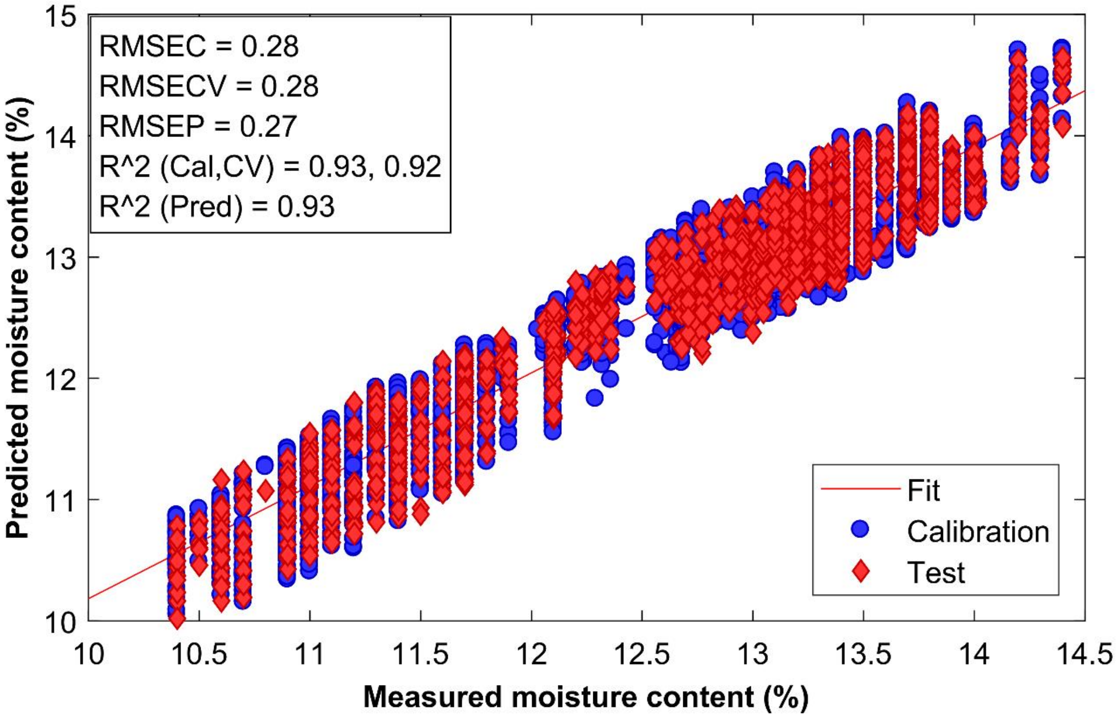

Figure 4.28 Measured vs. predicted moisture content for PLS regression model with SNV pre-treatment for the SK triticale data set for model 3 (Table 4.8) using 14 LV's and a calibration set of 3296 and validation set of 1325 SK's.

The combined data set for SK moisture content prediction using robust-PLS and spectral pre-treatment with SNV and DT an RMSEP of 0.51% and an R^2_P of 0.79 was obtained (Table 4.7). Applying GLS to the same SK data set, less LV's (9) were used resulting in an RMSEP of 0.53 with an R^2_P of 0.78 (Table 4.7). When outliers were removed manually from the combined SK data set and spectral pre-treatment with SNV an RMSEP of 0.23% with an R^2_P of 0.95 was achieved (Table 4.8). Applying GLS to the data set resulted in using 8 LV's to obtain an RMSEP of 0.29% and an R^2_P of 0.92 (Table 4.8).

The combined SNV pre-treated spectra shows no significant spectral outliers or extreme absorbance values (Fig. 4.29a). The VIP scores shown in Fig. 4.29b indicates that the first overtone of water (1450 nm) and the NH combination bands (1930 nm) are of significant value for moisture content prediction. The LV's shown in Figure 4.29c for LV1, LV2 and LV17, the first overtone of water is also indicated as being significant for moisture content prediction.

The RMSEC, RMSECV and RMSEP for increasing LV's is shown, indicating escarpments at 3 and 7 LV's and a steady decrease in error with no model overfitting being prevalent at higher LV's (Fig. 4.30). The predicted vs measured values for the combined SK data set, pre-treated with SNV (Table 4.8), shows no vertical residuals which can be considered outliers and a good spread around the model leverage (Fig. 4.30).

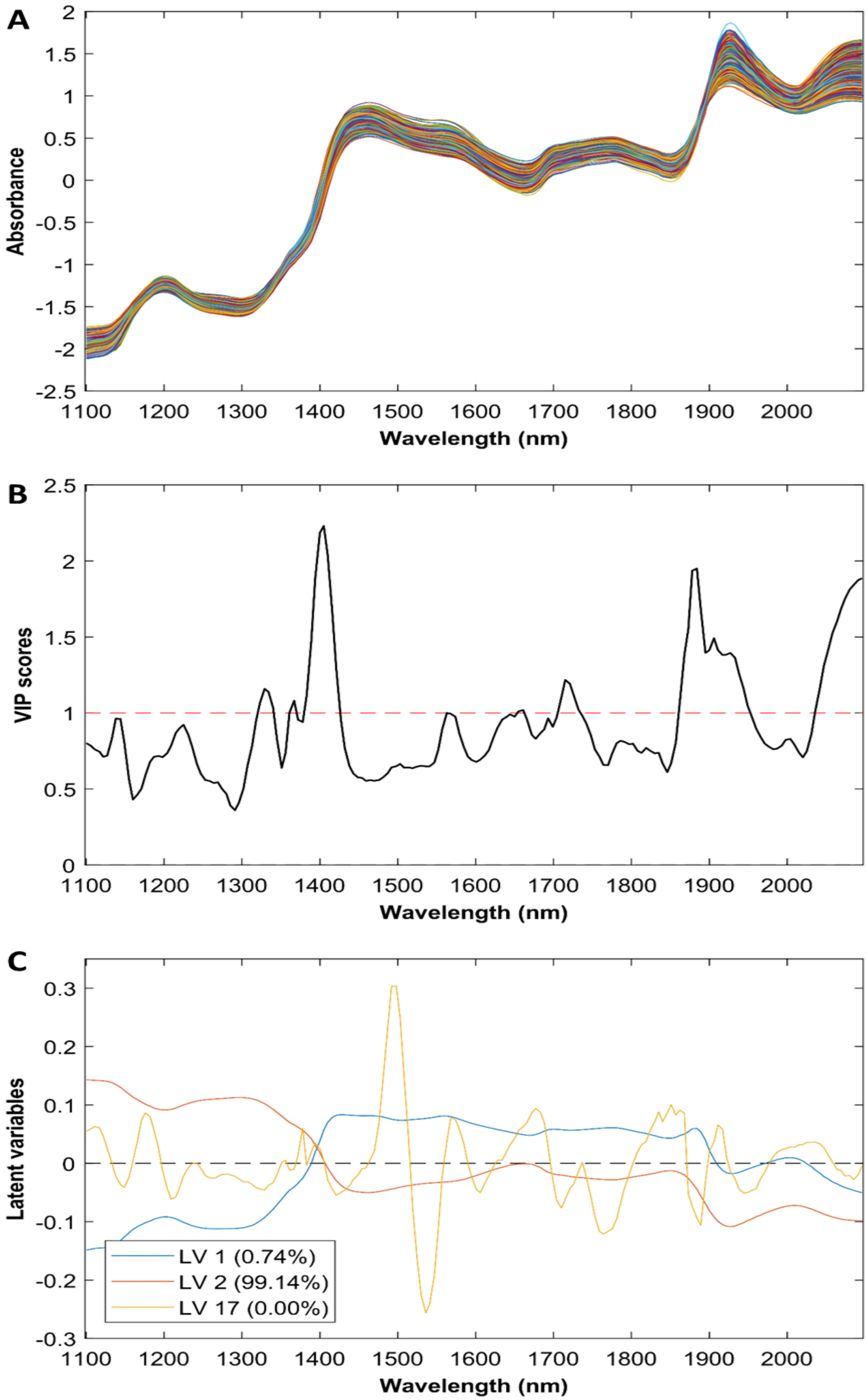

Figure 4.29 (a) Pre-processed spectra of the combined SK data set for moisture content prediction model 3 (Table 4.8), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 17.

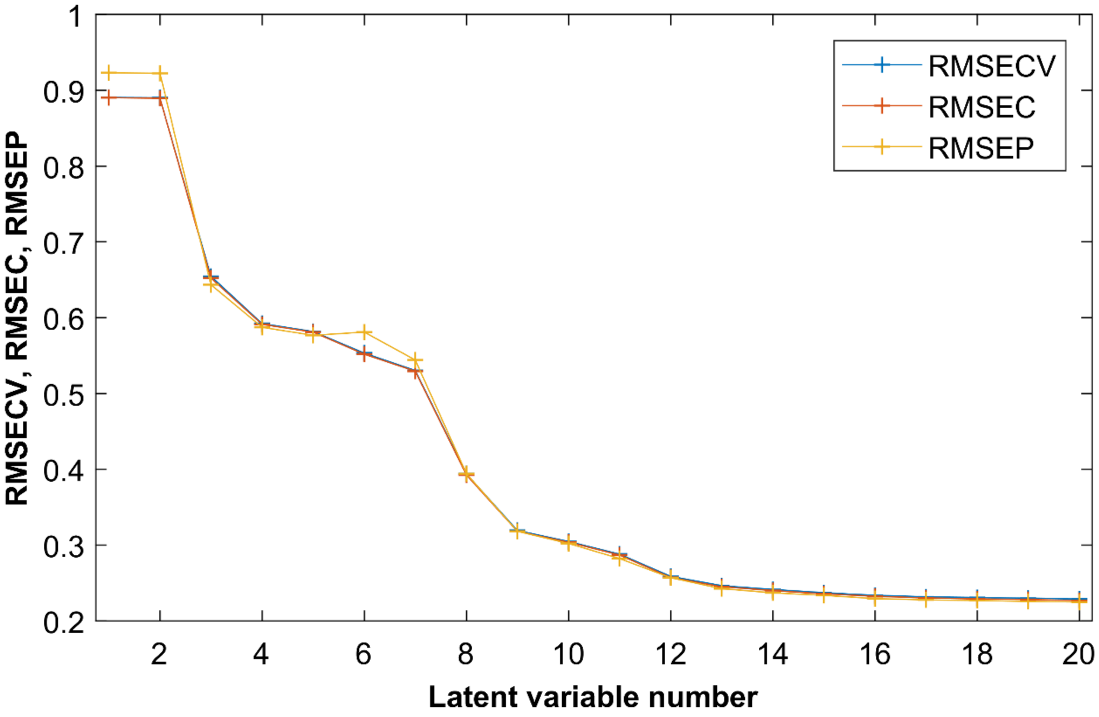

Figure 4.30. Latent variables vs standard error of cross-validation, -calibration and -prediction error for SK combined wheat and triticale moisture content prediction model 3 (Table 4.8).

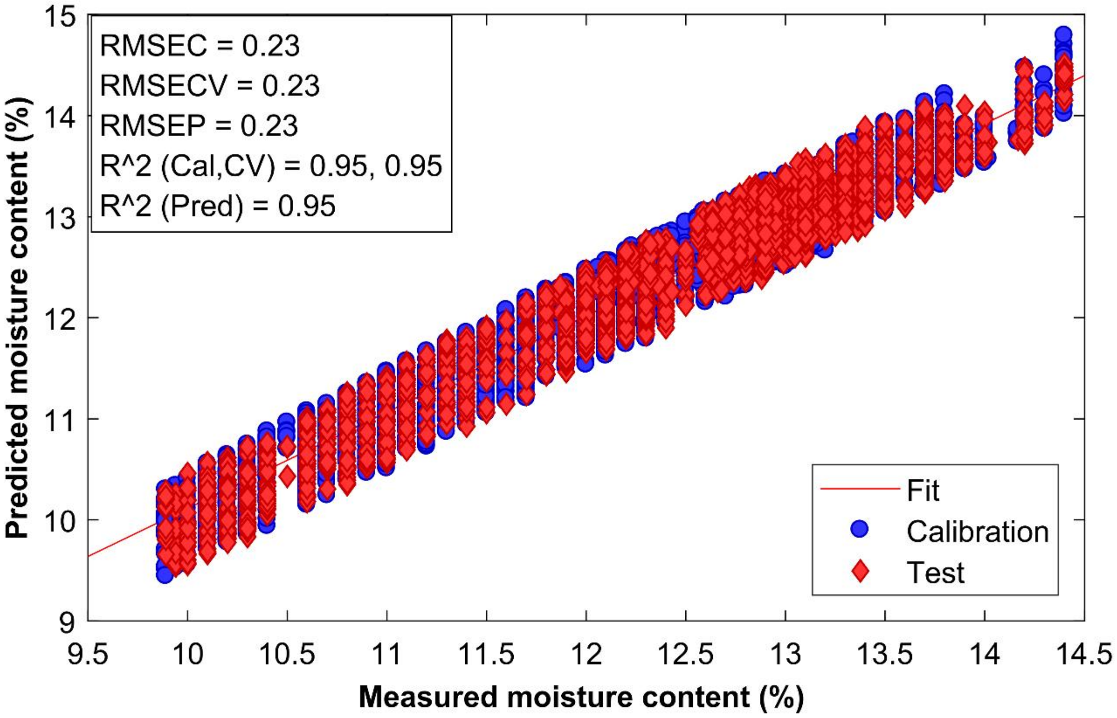

Figure 4.31 Measured vs. predicted moisture content for PLS regression model with SNV pre-treatment for the SK combined wheat and triticale data set (Table 4.8) using 17 LV's and a calibration set of 6722 and validation set of 2894 SK's.

4.5 Single kernel hardness

Single kernel PLS-regression models were calculated for wheat, triticale and the combined wheat and triticale kernel hardness data sets. The models were built by removing outliers manually, evaluating different spectral pre-treatment methods and using more or less LV's (Table 4.9). Using 2nd derivative and GLS as spectral pre-treatment on the SK wheat data set resulted in an RMSEP of 3.64 and an R^2_P of 0.62 with 3 LV's (Table 4.9). Applying the same spectral pre-treatment (2nd derivative and GLS) to the SK triticale data set resulted in an RMSEP of 2.09 and an R^2_P of 0.72 with 4 LV's. And for the combined wheat and triticale data set an RMSEP of 2.95 and an R^2_P of 0.69 with 4 LV's (Table 4.9).

The pre-treated spectra, VIP scores and latent variables for the SK wheat hardness data set (Table 4.9) are shown in Figure 4.32. The 2nd derivative and GLS pre-treated SK wheat spectra (Fig. 4.32a) give an indication that the first overtone of water (1450 nm) and protein absorbance region (1930 nm) have an impact towards kernel hardness prediction. The VIP scores plot (Fig. 4.32b) and the latent variable plot (Fig. 4.32c) also highlights that the regions of importance for hardness prediction are the C-H combination bands at 1370-1390 nm, the first overtone of water and protein absorbance regions at 1430-1500 nm. The protein absorbance region at 1430 nm is highlighted as being more important than the first overtone of water in the VIP scores plot. In addition, it also indicates that the C-O, 2nd overtone at around 1900 nm contributes significant weight to wheat hardness prediction (Williams *et al.*, 2019).

The RMSECV, RMSEC and RMSEP vs. increase in LV's for the SK wheat hardness data set (Fig. 4.33) shows that after 4 LV's significant model overfitting is observed. It also indicates that calibration and cross-validation error are better than prediction error for the specific data set and spectral pre-treatment. The measured vs. predicted SK hardness (Fig. 4.34) shows that the coefficient of determination could be better resolved with smaller vertical residuals and more model leverage.

The pre-treated spectra, VIP scores and LV's for the SK triticale hardness data set (Table 4.9) are shown in Figure 4.35. The three complementing plots indicate that the hydrocarbon aliphatic and aromatic region (1370-1390 nm), the first overtone of water and the protein absorbance region

(1430-1500 nm) carry more weight than the cellulose and carboxylic acid region (1820-1900 nm) for hardness prediction (Williams *et al.*, 2019).

The RMSECV, RMSEC and RMSEP vs. increase in LV's for the SK triticale data set pre-treated with 2nd derivative and GLS (Fig. 4.36) shows overfitting after 5 LV's and lower calibration and cross-validation error compared to prediction error. The measured vs. predicted SK hardness plot (Fig. 4.37) shows that with smaller vertical residuals a better coefficient of determination will be obtained, the limitation however being the number of samples.

The pre-treated spectra, VIP scores and latent variables for the SK combined wheat and triticale hardness data set (Table 4.9) are shown in Figure 4.38. The three plots indicate that the hydrocarbon aliphatic and aromatic (1370-1390 nm), the first overtone of water (1410 nm), the cellulose and carboxylic acid (1800-1900 nm), the O-H combination bands for starch and water (1930-1960 nm) and the N-H combination bands (1980-2060 nm) contribute significantly towards hardness prediction for the combined data set.

The RMSEC, RMSECV and RMSEP vs. increase in LV's for the SK combined wheat and triticale data set pre-treated with 2nd derivative and GLS (Fig. 4.39) show a good decrease in error with increase in LV's. The plot indicates model overfitting after 6 LV's and that prediction accuracies were better than calibration and cross-validation with increase in LV's. The measured vs. predicted SK hardness shown in Figure 4.40 indicates that regression could be better with smaller vertical residuals and that the prediction set falls within the calibration set, a wider and more representative leverage can also increase model accuracy.

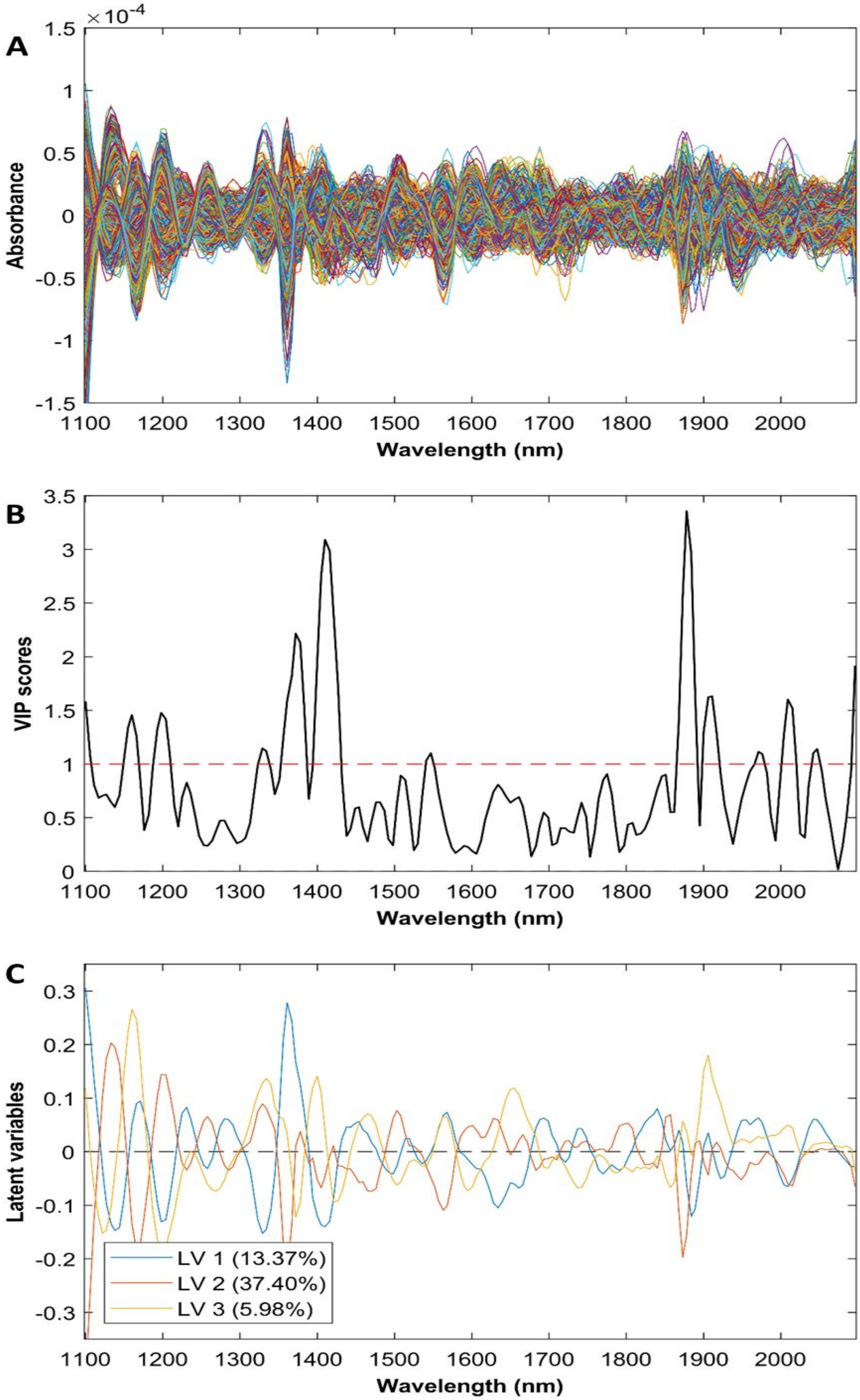

Figure 4.32 (a) Pre-processed spectra of the wheat SK data set for kernel hardness prediction model 2 (Table 4.9), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 3.

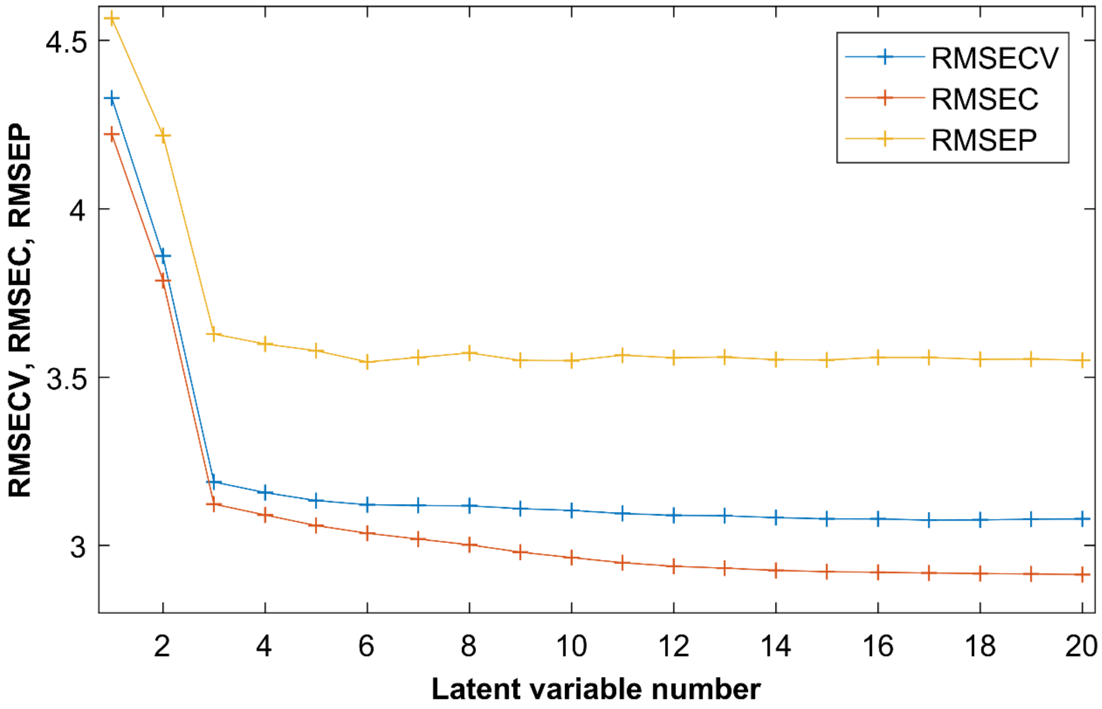

Figure 4.33 Latent variables vs standard error of cross-validation, -calibration and -prediction for SK wheat hardness prediction model 2 (Table 4.9).

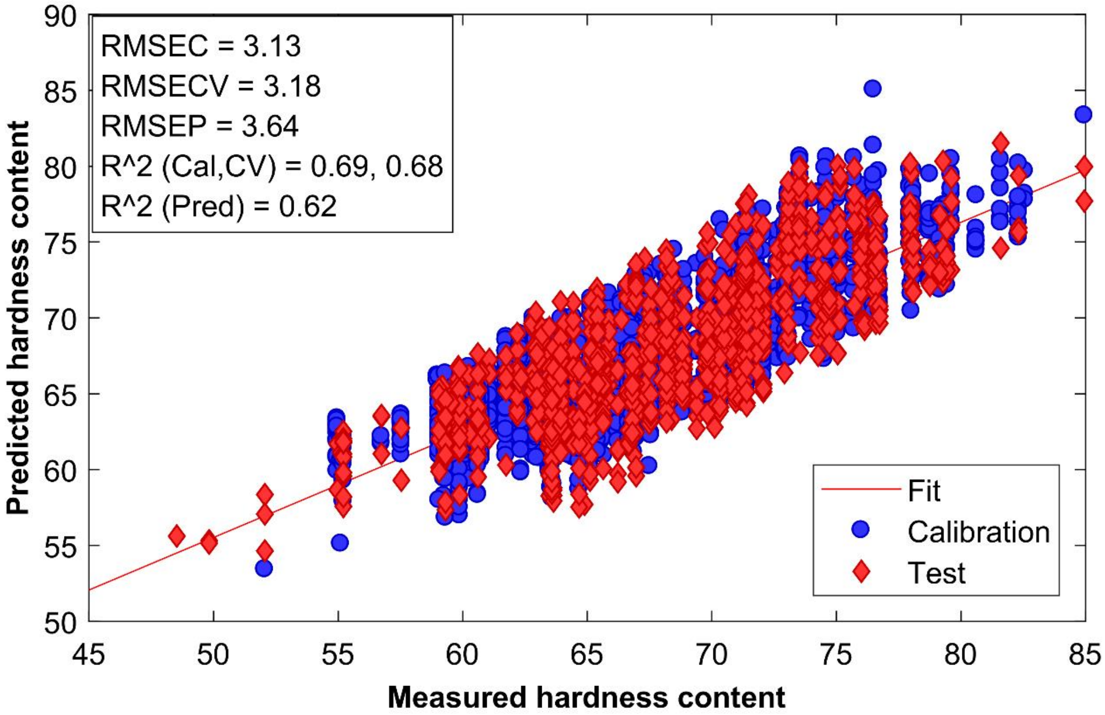

Figure 4.34 Measured vs. predicted hardness for PLS regression model 2 with GLS and SNV pre-treatment for the SK wheat data set (Table 4.9) with 3 LV's and a calibration set of 2362 and validation set of 886 SK's.

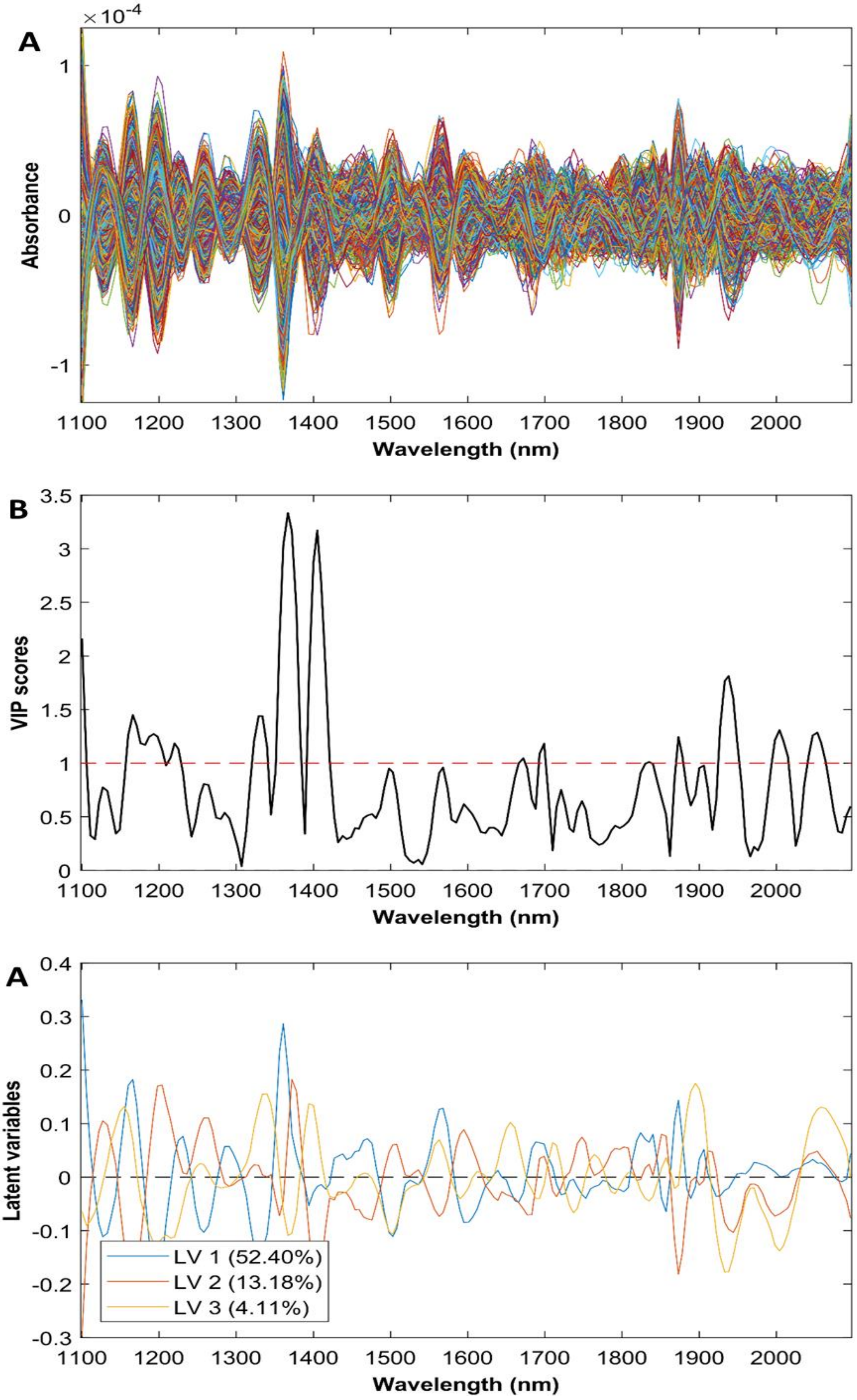

Figure 4.35 (a) Pre-processed spectra of the triticale SK data set for kernel hardness prediction model 2 (Table 4.9), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 3.

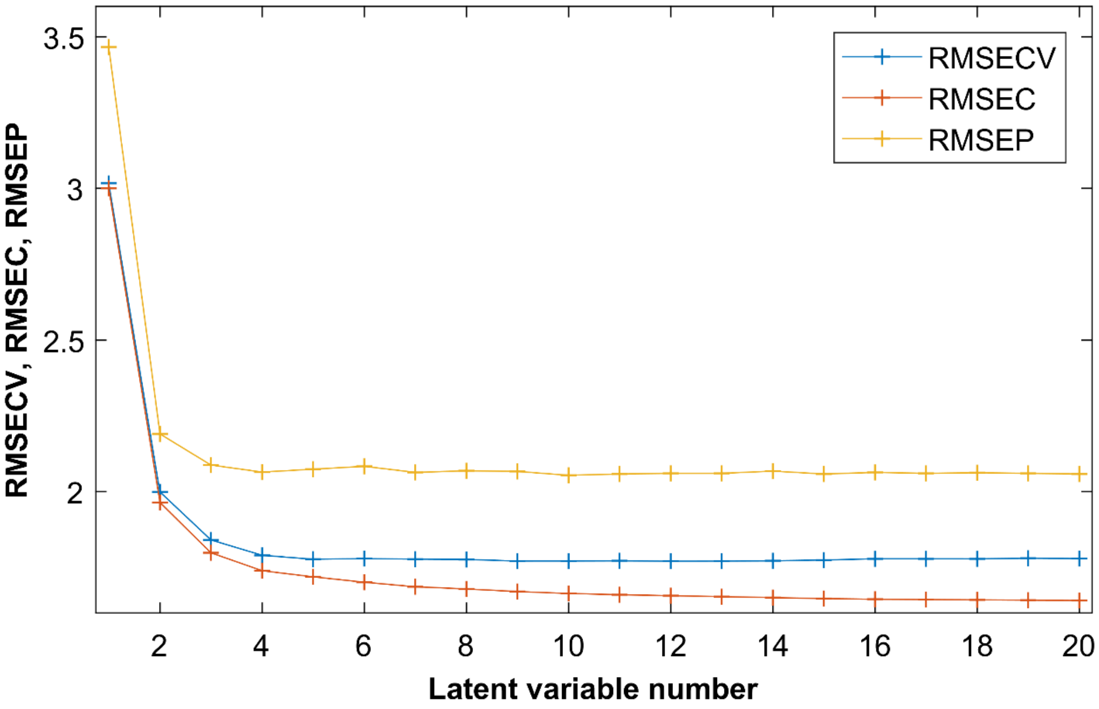

Figure 4.36 Latent variables vs standard error of cross-validation, -calibration and -prediction for hardness prediction model 2 (Table 4.9).

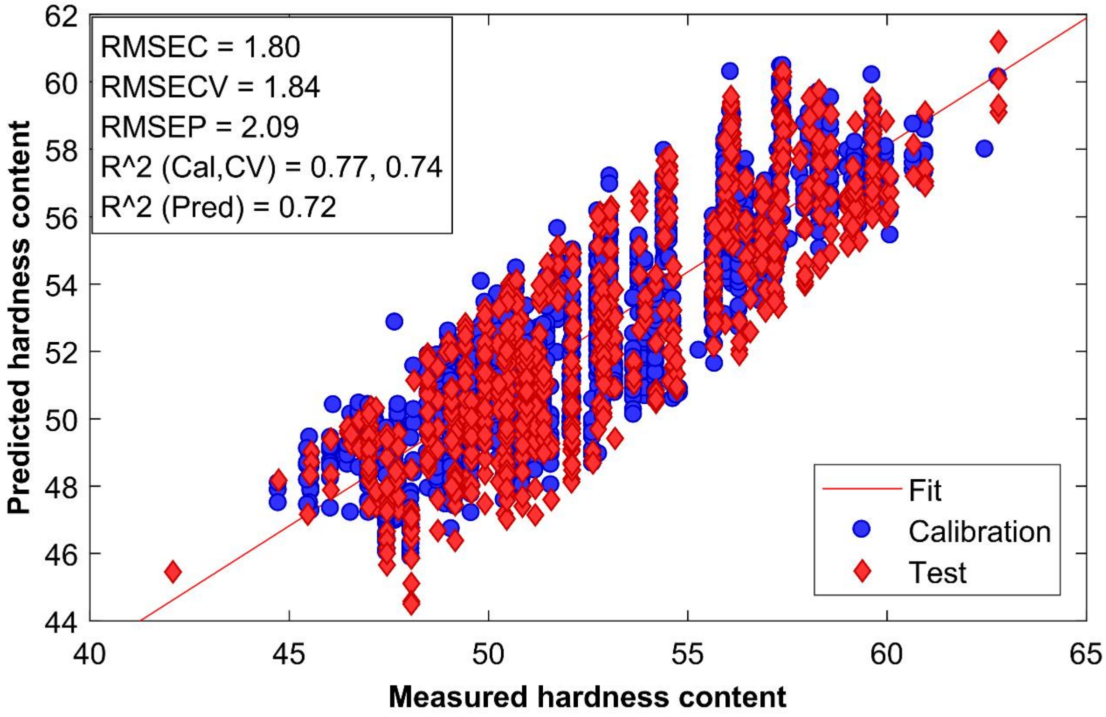

Figure 4.37 Measured vs. predicted hardness for PLS regression model with GLS and SNV spectral pre-treatment for the SK triticale data set model 2 (Table 4.9) using 4 LV's and a calibration set consisting of 1640 and validation of 658 SK's.

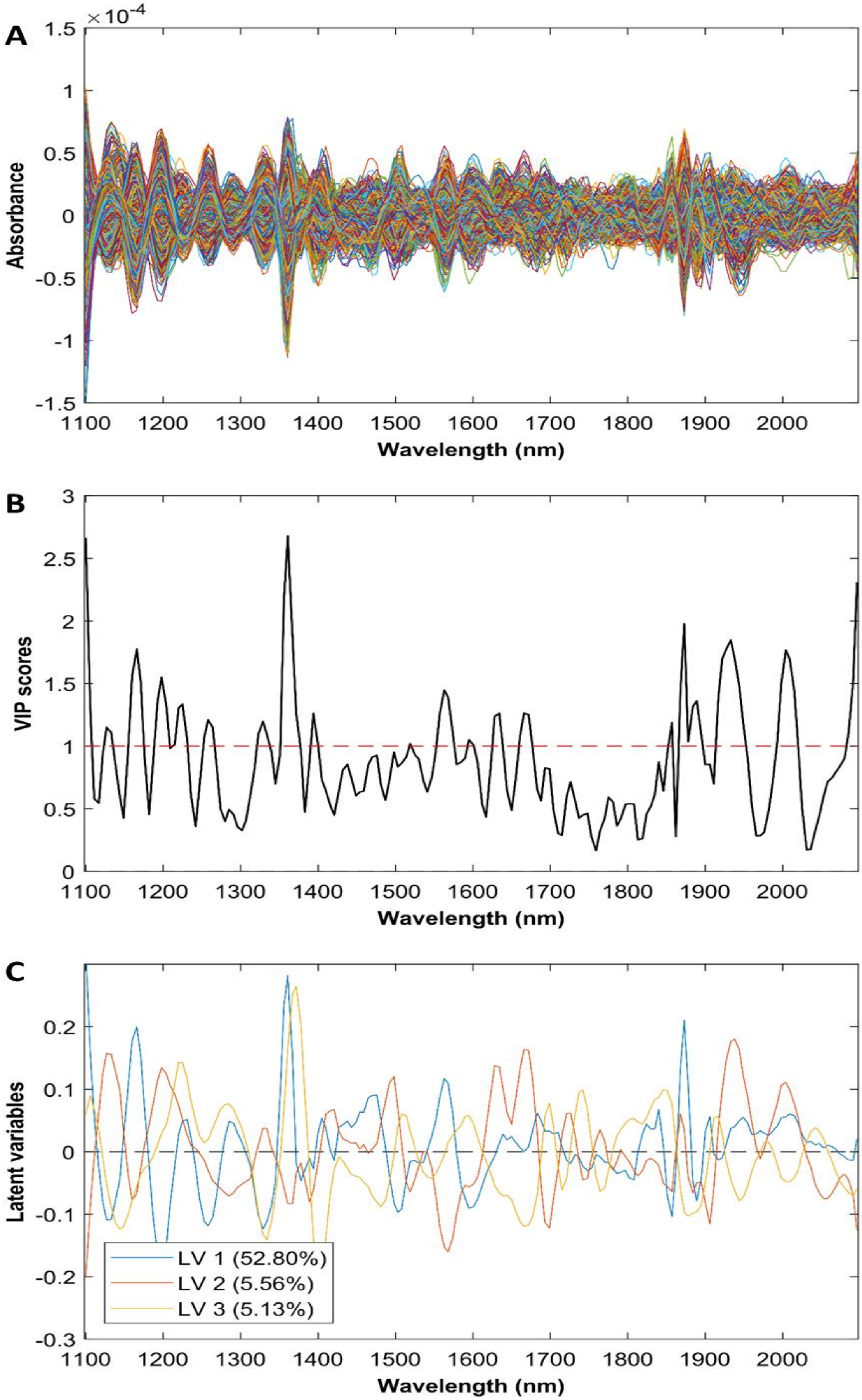

Figure 4.38 (a) Pre-processed of the combined SK data set for kernel hardness prediction model 2 (Table 4.9), (b) variable importance in projection and (c) LV's for LV 1, LV 2 and LV 3.

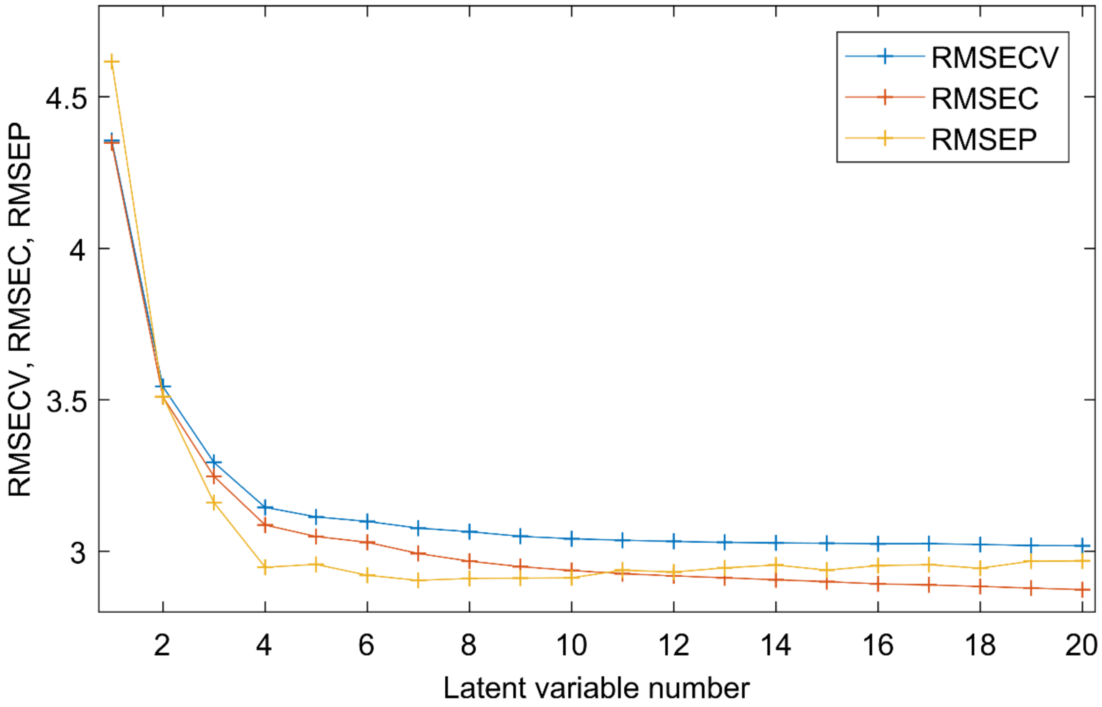

Figure 4.39. Latent variables vs standard error of cross-validation, -calibration and -prediction for hardness prediction model 2 (Table 4.9).

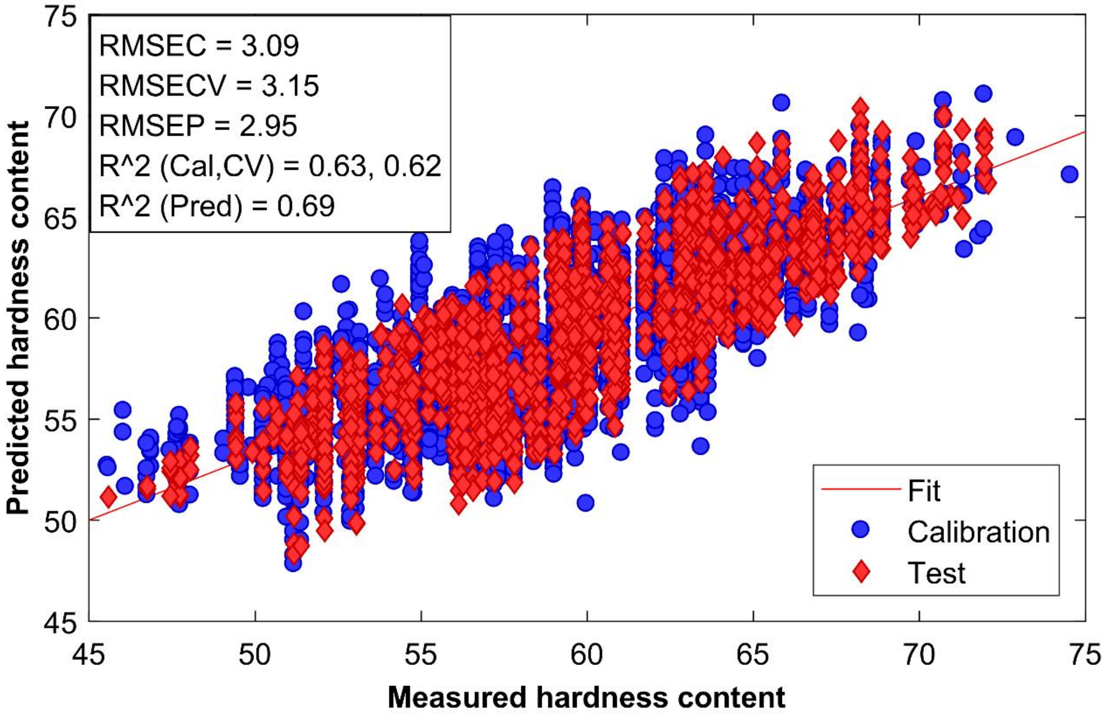

Figure 4.40 Measured vs. predicted hardness for PLS regression model with GLS and SNV spectral pre-treatment for the SK combined wheat and triticale data set model 2 (Table 4.9) using 4 LV's and a calibration set of 3259 and validation set of 1288 SK's.

Table 4.9 Calibration and validation statistics for predicted hardness PLS regression models for single kernel wheat, triticale and combined data sets using different pre-processing methods and removing outliers manually. The best prediction based on lowest RMSEP is indicated in bold

Wheat data set

#	**Preprocessing**	**XT**	**XV**	**LV**	**RMSEC**	**R^2_{C}**	**RMSECV**	**R^2_{CV}**	**RMSEP**	**R^2_{P}**
1	GLS (Y Gradient, alpha 0.0002)	2362	886	9	3.02	0.71	3.37	0.64	3.15	0.71
2	**2nd der (order: 3, window: 15 pt), GLS (Y Gradient, alpha 0.0002)**	**2362**	**886**	**3**	**3.12**	**0.69**	**3.18**	**0.68**	**3.64**	**0.62**
3	SNV, DT	2362	886	11	3.01	0.72	3.05	0.71	2.99	0.74
4	SNV	2362	886	11	3.02	0.71	3.06	0.71	2.98	0.74
5	DT	2362	886	12	3.09	0.70	3.15	0.69	3.14	0.71
6	OSC	2362	886	12	3.16	0.69	3.22	0.67	3.16	0.71
7	2nd der (order: 3, window: 15 pt)	2362	886	8	3.22	0.67	3.27	0.66	3.24	0.69
8	None	2362	886	11	3.14	0.69	3.19	0.68	3.11	0.72
CV	Venetian blind with 20 splits and 5 samples per split									

Triticale data set

#	Preprocessing	XT	XV	LV	RMSEC	R^2_{C}	RMSECV	R^2_{CV}	RMSEP	R^2_{P}
1	GLS (Y Gradient, alpha 0.0002)	1640	658	7	1.77	0.76	2.00	0.69	1.88	0.75
2	**2nd der (order: 3, window: 15 pt), GLS (Y Gradient, alpha 0.0002)**	**1640**	**658**	**4**	**1.80**	**0.75**	**1.84**	**0.74**	**2.09**	**0.72**
3	SNV, DT	1640	658	9	1.86	0.73	1.90	0.72	1.89	0.75
4	SNV	1640	658	8	1.91	0.72	1.94	0.71	1.90	0.74
5	DT	1640	658	8	1.83	0.74	1.86	0.73	1.77	0.78
6	OSC	1640	658	9	1.80	0.75	1.84	0.74	1.79	0.77
7	2nd der (order: 3, window: 15 pt)	1640	658	6	1.80	0.75	1.82	0.74	1.74	0.79
8	None	1640	658	9	1.82	0.74	1.86	0.73	1.79	0.77
CV	Venetian blind with 20 splits and 5 samples per split									

Combined data set

#	**Preprocessing**	**XT**	**XV**	**LV**	**RMSEC**	**R^2_{C}**	**RMSECV**	**R^2_{CV}**	**RMSEP**	**R^2_{P}**
1	GLS (Y Gradient, alpha 0.0002)	3250	1288	9	2.86	0.68	3.10	0.63	3.00	0.67
2	**2nd der (order: 3, window: 15 pt), GLS (Y Gradient, alpha 0.0002)**	**3250**	**1288**	**4**	**3.09**	**0.63**	**3.15**	**0.62**	**2.95**	**0.69**
3	SNV, DT	3250	1288	10	3.20	0.60	3.24	0.59	3.24	0.61
4	SNV	3250	1288	10	3.20	0.60	3.24	0.59	3.25	0.61
5	DT	3250	1288	10	3.19	0.60	3.23	0.59	3.25	0.61
6	OSC	3250	1288	8	2.97	0.66	3.00	0.65	3.00	0.67
7	2nd der (order: 3, window: 15 pt)	3250	1288	12	3.16	0.61	3.22	0.60	3.23	0.62
8	None	3250	1288	9	2.97	0.66	2.99	0.65	2.99	0.67
CV	Venetian blind with 20 splits and 5 samples per split									

4.6 Single kernel protein content independent test set

The SK wheat, triticale and combined data set models for prediction of protein content percentage and following the method of removing vertical outliers and bad leverage point using robust-PLS (Table 4.5) and manually using only conventional PLS (Table 4.6) were tested against an independent test set. Results are shown in Table 4.10 and Figures 4.41 (wheat), 4.42 (triticale) and 4.43 (combined wheat and triticale data set). The prediction accuracies for the SK independent test set was not as good as the validation sets shown in Tables 4.5 and 4.6. robust-PLS prediction accuracy, however, was slightly better than conventional PLS, indicating that the models were indeed more robust towards prediction of new SK's previously unseen by the model. Predicted protein content vs. residuals shown in Figures 4.41, 4.42 and 4.43 for both conventional PLS and robust-PLS indicate a wide scattering of vertical residuals around the calibrated models protein reference ranges. The robust-PLS predictions are more centroid around the mean of the reference values to which the models were calibrated for. Both methods signify the need for a greater degree of variance in model calibration reference values to achieve better prediction accuracy.

Table 4.10 RMSEP for SK independent test set of wheat, triticale, and the combined data set tested against the best prediction models for the robust-PLS and conventional PLS methods (Tables 4.5 and 4.6)

Data set	Test count	Robust-PLS RMSEP	PLS RMSEP
Wheat	76	2.70	2.81
Triticale	73	1.95	1.92
Combined	149	2.37	2.40

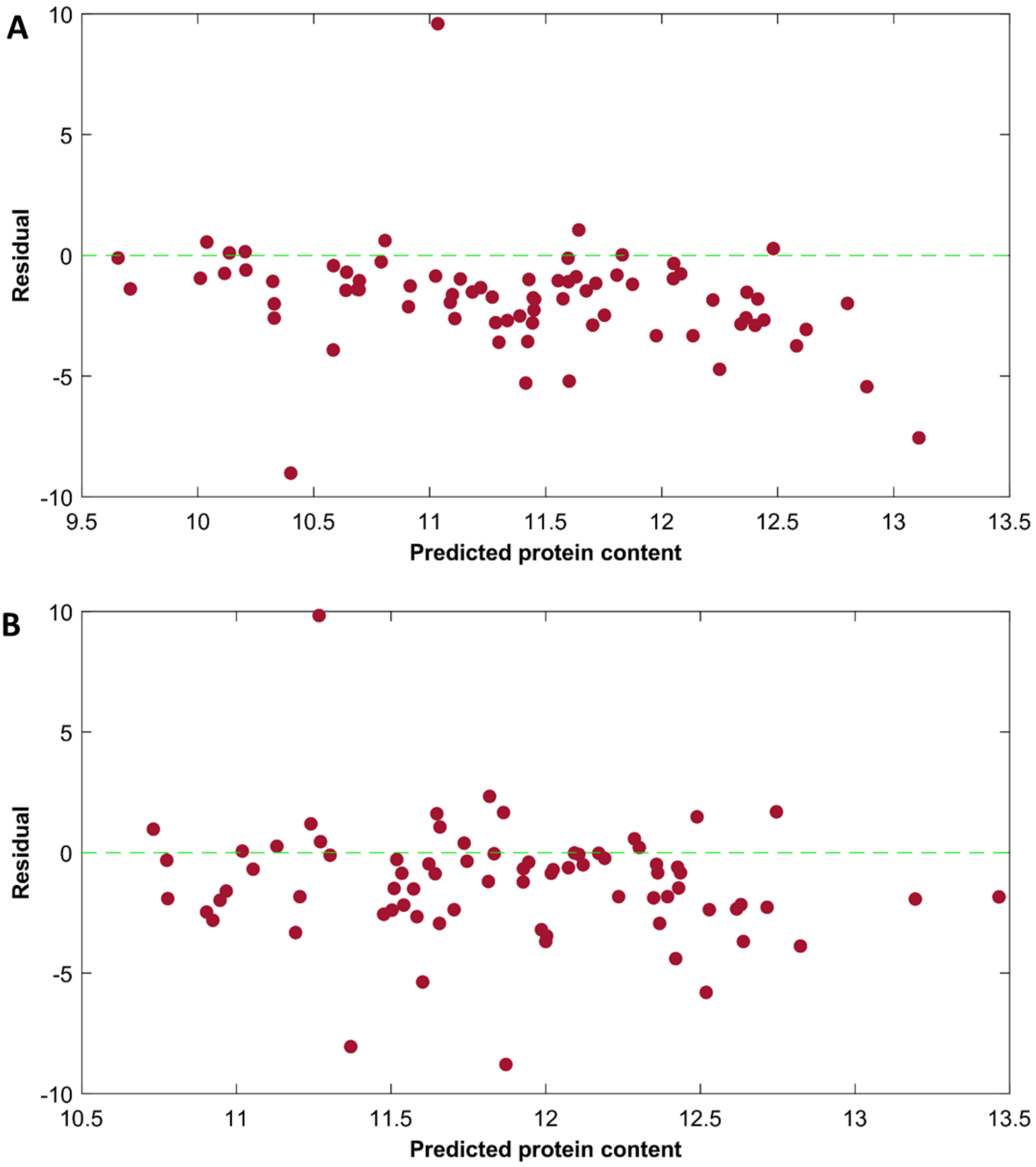

Figure 4.41 Predicted protein content vs. residual reference values for wheat SK test set, obtained for conventional (a) PLS and (b) robust-PLS methods (Tables 4.5 and 4.6).

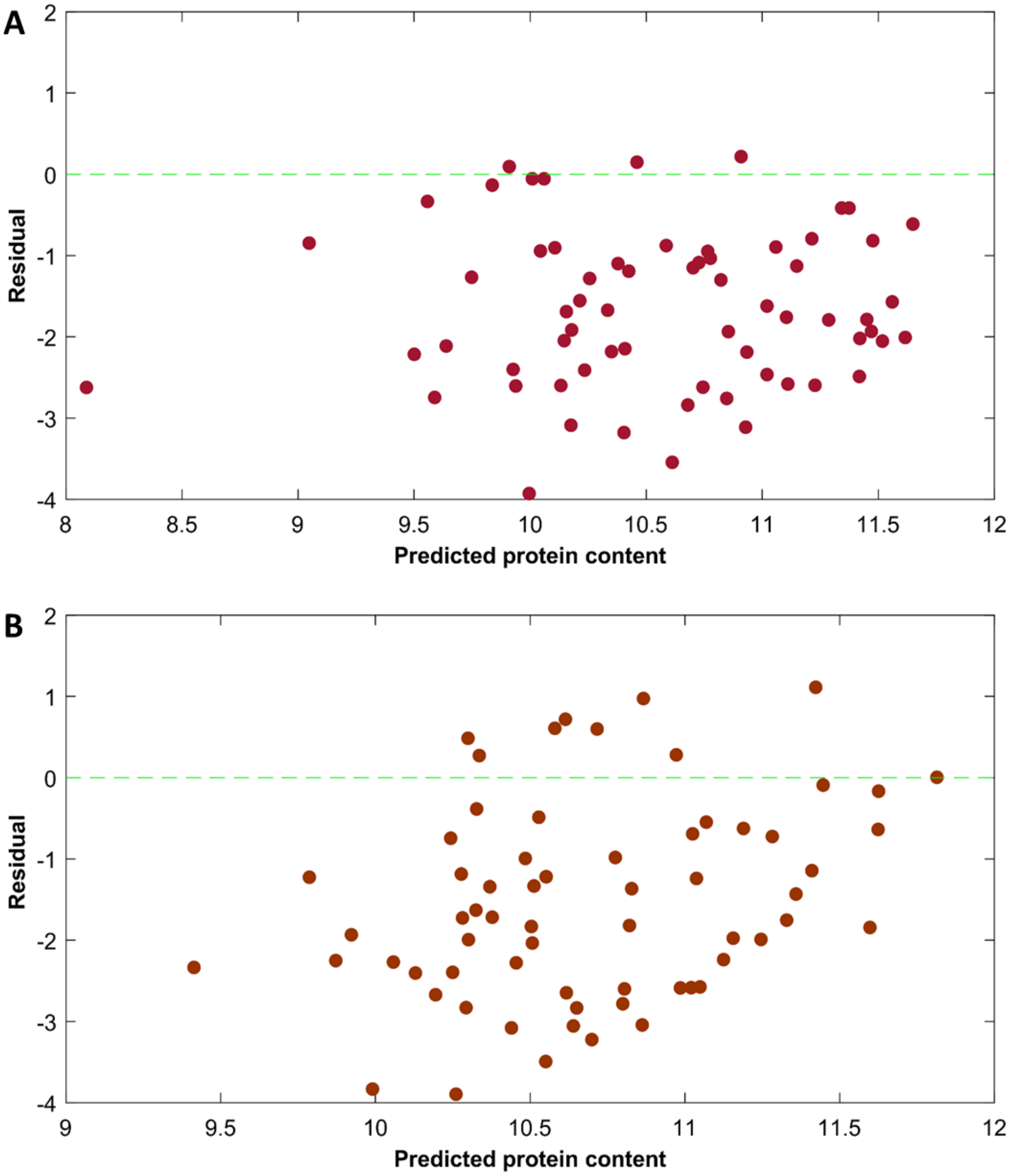

Figure 4.42 Predicted protein content vs. residual reference values for triticale SK test set, obtained for conventional (a) PLS and (b) robust-PLS methods (Tables 4.5 and 4.6).

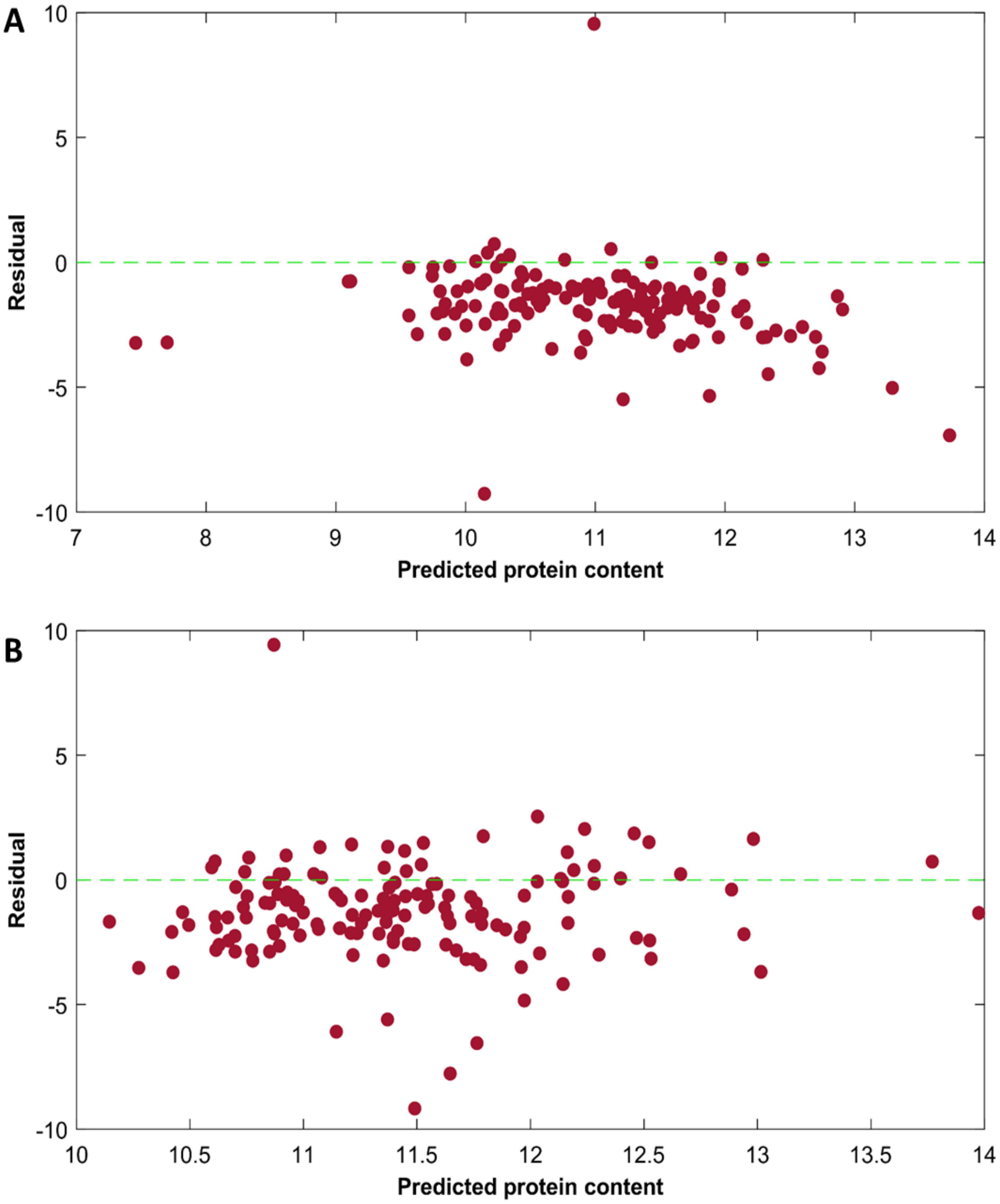

Figure 4.43 Predicted protein content vs. residual reference values for the combined SK test set, obtained for conventional (a) PLS and (b) robust-PLS (Tables 4.5 and 4.6).

Single kernel results for the prediction of protein and moisture content and kernel hardness proved to be good. Optimal results were obtained when outliers were removed manually compared to the robust-PLS method. This could be because manual outlier removal based on Y-residuals is more selective than that of robust-PLS which utilises the same outlier removal criteria as what was tried to achieve manually. However as also highlighted in the method development study by Hubert and Van den Branden (2003), the SIMPLS method fares better than the robust-SIMPLS method when only the calibration set and validation set are considered, but when an independent test set is subjected to the model the inverse is true. The same results were apparent for this study as the robust-PLS method obtained better prediction accuracy compared to the conventional PLS method for the independent test set. When GLS weighting was used as the pre-treatment method, favourable results were obtained for all models with the highlight being the low number of LV's being used to give similar results as the conventional spectral pre-treatment.

The effectiveness of SNV and DT towards prediction accuracy on SK's can be attributed to the phenotypical nature of wheat and triticale kernels which are prone to spectral scattering effects. Another aspect attributed to good spectra, was by using a NIR-HSI microscope lens that shortened the light path length towards the sensor. The shortened path to the sensor inherently was responsible for spatial filtering of the reflected light, reducing Lorenz-Mei and Rayleigh (scattering around spherical objects) scattering effects from the SK's (Lu *et al.*, 2006). Another pre-treatment technique that showed a good reduction in the amount of LV's used was GLS weighting with a Y block gradient. GLS removed reference data that did not specifically match that of the spectra systematically compared to similar reference data.

Triticale protein and moisture content and also kernel hardness models are comparable to the studies performed for SK wheat analysis using conventional NIR-spectroscopy and spectral imaging, highlighted in Chapter 2. As no other studies of such a nature using triticale as the subject of NIR-HSI analysis have been performed, a real comparison cannot be made and this study ultimately sets the benchmark. The benchmark was also set by this study for combining the spectra of two cereal species (wheat and triticale) into one data set and building NIR-HSI PLS-R SK protein and moisture content and also kernel hardness prediction models. The combined data set models showed good performance with a large sample set for rapid and non-invasive prediction of SK's based on both their protein and moisture content. The advantages of NIR-HSI as a rapid tool for single kernel analysis has been highlighted over that of bulk analysis methods, and also gives the ability to visualise the composition of a cereal grain and to go as far as to analyse on a pixel wise approach (Fox and Manley, 2014). Variance between kernels within a sample subset can easily be determined using SK predictions and it gives an indication of the uniformity of the sample subset.

Surprisingly SK wheat hardness determination using NIR-HSI has only vaguely been explored by Erkinbaev *et al*. (2019). The authors show that with a calibration set of 130 kernels and a test set of 30 kernels and by using conventional PLS-R and artificial neural network (ANN) models it was possible to quantify kernel hardness. The results of Erkinbaev *et al*. (2019) compare favourably with that of this study – with only for the regression coefficient being better in the current study. This could be due to the much larger data set of this study having more vertical residuals and bad leverage points compared to the smaller sample set used in the study of Erkinbaev et al. (2019). The over explanation of spectral data compared to reference data is another aspect which contributes toward models which could be better optimised, i.e. not enough SK reference data was collected compared to spectral data. Prediction accuracy for kernel hardness could also perhaps be increased by using advanced neural network techniques.

The ratio of performance to deviation (RPD) vs. R^2 is plotted (Fig. 4.44) for the wheat and triticale SK prediction models obtained through SIMPLS and RSIMPLS algorithms. The plot indicates models with a regression statistic above 0.75 and RPD above 2 (outlined in red) are deemed suitable for further model development. Prediction models falling under this threshold can be re-evaluated by using an independent test set, if no increase in R^2 vs. RPD is noted these models should not be continued with.

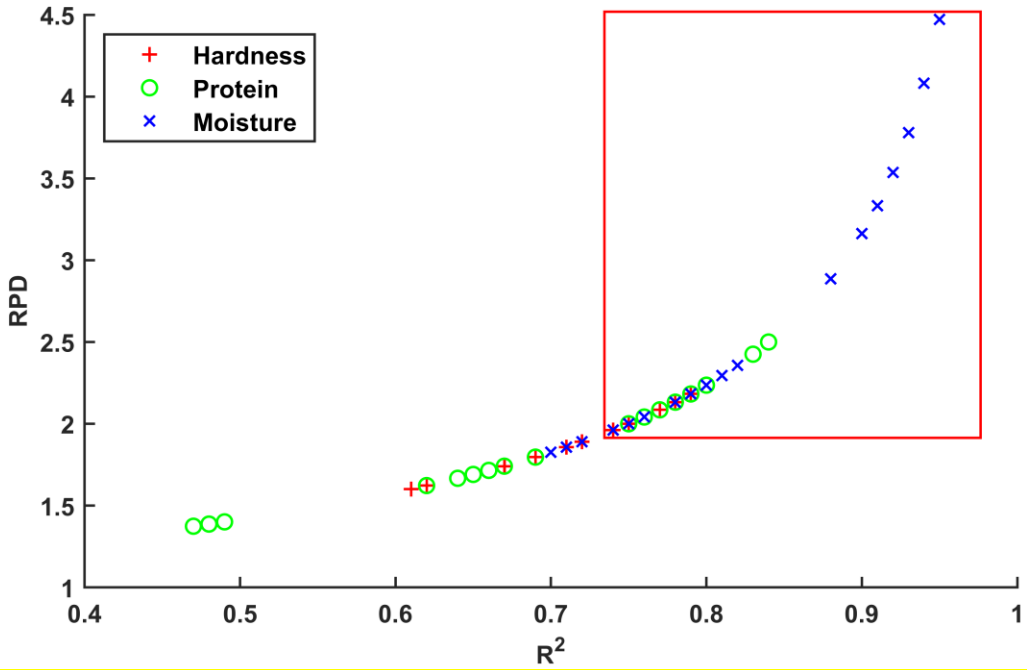

Figure 4.44 R^2-values plotted against RPD statistics for the SK wheat, triticale and combined data set models.

4.8 Conclusion

NIR-HSI PLSR models were reported on and it was able to show that fast, non-invasive and non-biased results could be obtained for the prediction of protein, moisture and kernel hardness content on a SK level for wheat, triticale and for the combined wheat and triticale data set. It was also shown to be possible to obtain good calibrations with comparable results to other studies by using bulk kernel image analysis for model building. A protein range of 7.11-14.66%, a moisture range of 9.89-14.40% and a hardness range of 14.69-84.96 was significant to produce good reference values for NIR-HSI PLS-R modelling. It was also possible to obtain R^2-values of above 0.75 and RMSEP values below 0.50 were obtainable for both the bulk and SK wheat, triticale and the combined wheat and triticale data sets. The study indicated that a large number of samples needs to be supplemented with seasonal data and also with samples that differ in growing origin to truly have a robust and well-rounded model. And finally by using advance pre-treatment techniques such as GLS it was possible to reduce LV's to obtain more robust and less complex models.

4.9 References

Delwiche, S.R. & Hruschka, W.R. (2000). Protein content of bulk wheat from near-infrared reflectance of individual kernels. *Cereal Chemistry*, **77**, 86–88.

Dowell, F.E., Maghirang, E.B., Xie, F., Lookhart, G.L., Pierce, R.O., Seabourn, B.W., Bean, S.R., Wilson, J.D. & Chung, O.K. (2006). Predicting wheat quality characteristics and functionality using near-infrared spectroscopy. *Cereal Chemistry Journal*, **83**, 529–536.

Erkinbaev, C., Derksen, K. & Paliwal, J. (2019). Single kernel wheat hardness estimation using near infrared hyperspectral imaging. *Infrared Physics and Technology*, **98**, 250–255.

Fox, G. & Manley, M. (2014). Applications of single kernel conventional and hyperspectral imaging near infrared spectroscopy in cereals. *Journal of the Science of Food and Agriculture*, **94**, 174–179.

Hubert, M. & Branden, K. Vanden. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*, **17**, 537–549.

Igne, B., Gibson, L.R., Rippke, G.R., Schwarte, A. & Hurburgh, C.R. (2007). Triticale moisture and protein content prediction by near-infrared spectroscopy (NIRS). *Cereal Chemistry*, **84**, 328–330.

Lu, B., Morgan, S.P., Crowe, J.A. & Stockford, I.M. (2006). Comparison of methods for reducing the effects of scattering in spectrophotometry. *Applied Spectroscopy*, **60**, 1157–1166.

Maghirang, E.B. & Dowell, F.E. (2003). Hardness measurement of bulk wheat by single-kernel visible and near-infrared reflectance spectroscopy. *Cereal Chemistry*, **80**, 316–322.

Mahesh, S., Jayas, D.S., Paliwal, J. & White, N.D.G. (2014a). Comparison of partial least squares regression (PLSR) and principal components regression (PCR) methods for protein and hardness predictions using the near-infrared (NIR) hyperspectral images of bulk samples of canadian wheat. *Food and Bioprocess Technology*, **8**, 31–40.

Manley, M., McGoverin, C.M., Snyders, F., Muller, N., Botes, W.C. & Fox, G.P. (2013). Prediction of triticale grain quality properties, based on both chemical and indirectly measured reference methods, using near-infrared spectroscopy. *Cereal Chemistry*, **90**, 540–545.

Manley, M., Zyl, L. Van & Osborne, B.G. (2002). Using fourier transform near infrared spectroscopy in determining kernel hardness, protein and moisture content of whole wheat flour. *Journal of Near Infrared Spectroscopy*, **10**, 71-76.

Williams, P., Antoniszyn, J. & Manley, M. (2019). Near-infrared Technology: Getting the best out of light. *Near-infrared Technology: Getting the best out of light*. AFRICAN SUN MeDIA.

Williams, P.C., Norris, K.H. & Sobering, D.C. (1985). Determination of protein and moisture in wheat and barley by near-infrared transmission. *J. Agric. Food Chem*.

Chapter 5: General discussion and conclusions

Wheat (*Triticum aestivum*) and triticale (× Triticosecale sp. Wittmack ex A. Camus 1927) grain is used for human and animal consumption, with wheat being used as an important human nutritional source and triticale mostly being used as animal feed. Quality measurements for these grains include amongst others the determination of protein and moisture content (wheat and triticale) and in some cases kernel hardness (wheat) in order to select lines in breeding programmes, appropriate for end use and the sale price. Conventional methods for the quantification of these quality properties include destructive techniques such as Dumas combustion (protein content), air oven drying (moisture content) and the Single Kernel Characterisation System (SKCS; kernel hardness). Near-infrared (NIR) spectroscopy is used as a rapid alternative method which offers the advantage of being non-invasive and non-destructive. Conventional NIR spectroscopy, is mainly applied to bulk samples and provides predictions based on an average spectrum. More recently NIR hyperspectral imaging has been considered with the added advantage of a spatial dimension. This enables the analysis of multiple single kernels simultaneously and rapidly, but with the option of predicting properties based on individual kernels. The spatial dimension in addition provides the potential of determination distribution of chemical components within each kernel. NIR hyperspectral imaging (NIR-HSI) is thus an analytical tool with a high spectral as well as spatial resolution to gather data rapidly and non-invasively. This study aimed to develop wheat and triticale NIR-HSI partial least squares regression models to accurately predict protein and moisture content and kernel hardness. Models were developed for bulk samples as well as on a single kernel (SK) basis.

In this study 180 wheat and 177 triticale samples, originating from 3 growing regions in South Africa, were used to obtain images (1100-2100 nm) for the bulk and SK samples sets In addition to the bulk sample images (180 wheat; 177 triticale), this resulted in a data set of 7020 wheat and 6903 triticale SK images. Partial least squares regression (PLS-R) models were subsequently developed and tested for the prediction of kernel protein and moisture content and kernel hardness.

This study set the benchmark for NIR-HSI analysis on triticale, with the first PLS-R models being published for protein and moisture content prediction and also for kernel hardness prediction. This was achieved for a bulk sample subset approach and also for SK analysis of whole grains. The benchmark was also set by combining the spectral data of two types of grain (wheat and triticale) into one data set and to subsequently build PLS-R models for predicting protein and moisture content and also kernel hardness – once more for both a bulk approach and SK analysis of whole grains.

Prediction results with R^2-values of above 0.75 and RMSEP values below 0.50 were obtained for both the bulk and SK wheat, triticale and the combined wheat and triticale data sets using PLS-R. Two PLS-R methods were evaluated for SK analysis with a validation and independent test set, i.e. robust-PLS and conventional PLS-R. Both methods showed good results. The robust-PLS method coupled with generalised least squares (y-block grading) proved to be a good technique for complex data sets having unresolved points of leverage and an over explanation of spectral data compared to reference data. The best RMSEP results (protein content: 0.37-0.84%, moisture content: 0.23-0.57% and kernel hardness: 1.74-3.64) were obtained for the conventional PLS-R method when the validation set was considered. The independent test set for protein content prediction achieved better RMSEP results with the robust-PLS (1.95-2.37%) method, proving that the method did indeed have an effect on making the calibration data sets more robust.

Single kernel results for the prediction of protein and moisture content proved to be good. Optimal results were obtained when outliers were removed manually compared to the robust-PLS method. This could be because manual outlier removal based on Y-residuals is more selective than that of robust-PLS which utilises the same outlier removal criteria as when it was removed manually. As also shown by Hubert and Van den Branden, (2003), the SIMPLS method (PLS-R) fares better than the Robust-SIMPLS (robust-PLS) method when only the calibration set and validation set were considered, but when an independent test set was subjected to the model the inverse was true. When generalised least squares (GLS) weighting was used as the pre-treatment method, favourable results were obtained for all models with the highlight being the low number of latent variables (LV's) being used to give similar results as the conventional spectral pre-treatment.

SK protein content prediction of wheat kernels using hyperspectral imaging was attempted by Caporaso *et al.* (2018). The current study resulted in lower RMSEP values (0.37 vs. 0.944%). The study by Caporaso *et al*. (2018) shows a far greater error in prediction with a smaller calibration and also validation set being used compared to the current study. The SK wheat protein content range of 6.2-19.8% used by Caporaso *et al*. (2018) shows that the lower and higher regions were underrepresented. Caporaso *et al*. (2018) could potentially have achieved better calibration results using less LV's if an advanced spectral pre-treatment method such as GLS was applied as was the case for this study.

SK wheat hardness determination using NIR-HSI has been vaguely explored by Erkinbaev *et al.* (2019). The authors showed that with a calibration set of 130 kernels and a test set of 30 kernels and by using conventional PLS-R and artificial neural network (ANN) models it was possible to quantify kernel hardness. The ANN method performed much better than the PLS method and obtained an R^2-value of 0.90 and an RMSEP of 6.59 compared to the PLS result with (R^2 of 0.80; RMSEP of 12.90). The results of Erkinbaev *et al*. (2019) compared favourably with that of this study (RMSEP of 1.74-3.64), however with a higher R^2-value. The much larger dataset of the current study having much more vertical residuals and many leverage points compared to the much smaller sample set used by Erkinbaev et al. (2019). The over explanation of spectral data compared to reference data is another aspect which contributes toward models which could be better optimised.

The use of a wheat and triticale combined data set addressed a recommendation by Igne *et al.* (2007) to combine a triticale and wheat data set to obtain better prediction accuracy.

The authors found that using wheat models to predict triticale moisture and protein content was not suitable as the SEP was too large, however it was considered usable for screening. The current study showed good prediction accuracy for such a combined data set and it to be a potential powerful approach to predict protein and moisture content and kernel hardness of small grains using a single combined model.

The performances of the models in the current study could be improved with more seasonal variation and by inclusion of more samples with high and low protein and moisture contents and kernel hardness values for calibration development. Wavelength selection could also be performed at the 1400 nm, 1600 nm and 1800-2000 nm regions for the data set of this study, as these regions were highlighted as being of importance for quantifying protein and moisture content (*ca*. 1400; *ca*. 1800-2000 nm) and also kernel hardness (*ca*. 1600 nm). The identification of wavelengths of importance will allow for a multispectral approach, which requires less computational input, a more rapid output and a also a great reduction in instrument cost (Xiaobo *et al.*, 2010). In turn allowing for models to be integrated into systems mounted on unmanned aerial vehicles and also systems for rapid online screening of bulk grain at silos and mills. Unmanned aerial vehicles are not only useful for quantifying protein and moisture content and kernel hardness in the field, but also a technique that can be used to monitor crop health and traits (Hassan *et al.*, 2019).

An interesting approach to gain more detailed information from single kernels would be to perform proteomic work, coupling liquid chromatography-mass spectroscopy and NIR-HSI (Wesley *et al.*, 2008). A more detailed approach could be followed to identify hardness alleles within the seeds, being a decisive tool in early stage breeding programs (de Groot, 2019) and combining this with spectral imaging. Such methods can be expanded to other wheat properties. Omics could e.g. be used to identify the enzymes responsible for germination in wheat, allowing for a qualitative modelling approach using NIR-HSI to identify grain that has germinated in the field (Bose *et al.*, 2019). These omics approaches can be integrated to an in-line system which is set-up on a combine harvester, allowing for real time data collection of crop quality which also includes constant geo-referencing of harvest location (Risius *et al.*, 2015; Chen *et al.*, 2020)

Ultimately the findings of this study indicated that robust-PLS methods are a good option for resolving complex data sets, with good prediction accuracy possible. Another aspect of importance was to note that a very large data set does not necessarily coincide with a vast enough variance being included in the model – if the spectral and reference data do not carry the same weight. It was not a feasible option to analyse *ca*. 14 000 individual kernels for protein and moisture content to be used as reference results. Lastly the technique showcased that it can find its place within the industry and compete with conventional methods.

References

Bose, U., Broadbent, J.A., Byrne, K., Hasan, S., Howitt, C.A. & Colgrave, M.L. (2019). Optimisation of protein extraction for in-depth profiling of the cereal grain proteome. *Journal of Proteomics*, **197**, 23–33.

Caporaso, N., Whitworth, M.B. & Fisk, I.D. (2018). Protein content prediction in single wheat kernels using hyperspectral imaging. *Food Chemistry*, **240**, 32–42.

Chen, J., Lian, Y. & Li, Y. (2020). Real-time grain impurity sensing for rice combine harvesters using image processing and decision-tree algorithm. *Computers and Electronics in Agriculture*, **175**, 105591.

Erkinbaev, C., Derksen, K. & Paliwal, J. (2019). Single kernel wheat hardness estimation using near infrared hyperspectral imaging. *Infrared Physics and Technology*, **98**, 250–255.

Groot, G. de. (2019). *Genotyping South African wheat germplasm for hardness alleles*.

Hassan, M.A., Yang, M., Rasheed, A., Yang, G., Reynolds, M., Xia, X., Xiao, Y. & He, Z. (2019). A rapid monitoring of NDVI across the wheat growth cycle for grain yield prediction using a multi-spectral UAV platform. *Plant Science*, **282**, 95–103.

Hubert, M. & Branden, K. Vanden. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*, **17**, 537–549.

Igne, B., Gibson, L.R., Rippke, G.R., Schwarte, A. & Hurburgh, C.R. (2007). Triticale moisture and protein content prediction by near-infrared spectroscopy (NIRS). *Cereal Chemistry*, **84**, 328–330.

Risius, H., Hahn, J., Huth, M., Tölle, R. & Korte, H. (2015). In-line estimation of falling number using near-infrared diffuse reflectance spectroscopy on a combine harvester. *Precision Agriculture*, **16**, 261–274.

Wesley, I.J., Osborne, B.G., Larroque, O. & Bekes, F. (2008). Measurement of the protein composition of single wheat kernels using near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, **16**, 505–516.

Xiaobo, Z., Jiewen, Z., Povey, M.J.W., Holmes, M. & Hanpin, M. (2010). Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, **667**, 14-32.